DOCUMENT RESUME

ED 181 051 TM 009 953

TITLE	Proceedings of the Invitational Conference on Testing Problems (3rd, New York, New York, October 28, 1950).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	28 Oct 50
NOTE	112p.
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Accountants; *Achievement Tests; Aptitude Tests; College Entrance Examinations; Conference Reports; *Creative Ability; Creativity Tests; Dentists; Evaluation Criteria; Higher Education; Lawyers; *Occupational Tests; Physicians; *Professional Occupations; Psychometrics; Statistical Analysis; Test Construction; *Testing Problems; *Test Selection; Test Validity

ABSTRACT

The conference was organized into two sessions: validation of professional aptitude batteries and criteria for the evaluation of achievement tests. Panelists and their topics were: Tests for Accounting, Arthur E. Traxler and Robert Jacobs; Tests for Law, A. Pemberton Johnson; Tests for Dentistry, Shailer Peterson; Tests for Medicine, John M. Stalnaker; From the Point of View of the Test Editor, Frederick B. Davis; From the Point of View of the Subject Matter Specialist, Joseph J. Schwab; From the Point of View of Their Internal Statistics, John B. Carroll; and From the Point of View of Their External Statistical Relationships, Harold Gulliksen. An address, Creative Talent, was delivered by L.L. Thurstone. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181051

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Proceedings

ETS

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

1950

TM009 953

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Educational Testing Service

EDUCATIONAL TESTING SERVICE

BOARD OF TRUSTEES

Raymond B. Allen, *Chairman*

Joseph W. Barker	Herold C. Hunt
Oliver C. Carmichael, *ex officio*	Katharine E. McBride, *ex officio*
Charles W. Cole	Thomas R. McConnell
James B. Conant	Lester W. Nelson
Henry H. Hill	Edward S. Noyes

George F. Zook, *ex officio*

OFFICERS

Henry Chauncey, *President*
Richard H. Sullivan, *Vice President and Treasurer*
William W. Turnbull, *Vice President*
Jack K. Rimalover, *Secretary*
Catherine G. Sharp, *Assistant Secretary*
Robert F. Kolkebeck, *Assistant Treasurer*

COPYRIGHT, 1951, EDUCATIONAL TESTING SERVICE,
20 NASSAU STREET, PRINCETON, N.J.
PRINTED IN THE UNITED STATES OF AMERICA

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

October 28, 1950

ROBERT L. THORNDIKE, *Chairman*

¶ Validation of professional aptitude batteries.

¶ Creative talent.

¶ Criteria for the evaluation of achievement tests.

◆

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY LOS ANGELES, CALIFORNIA

FOREWORD

Over the years the Invitational Conference on Testing Problems has grown in size and general interest. Not only have the papers proved stimulating, but the occasion has become a friendly gathering of those professionally interested in testing. The 1950 Conference maintained the high level established by its predecessors. Because of the significance of the papers, those who attended the meeting and others who were unable to attend will, I am sure, wish to have a full record of the proceedings, which Educational Testing Service is glad to publish.

I would like to express my appreciation to the participants, especially Dr. Thurstone, whose paper on creative ability will be a landmark in the attempt to develop measures of that elusive quality. My particular thanks go also to Robert L. Thorndike, whose capable chairmanship from the planning stage through the meeting itself made possible a most successful conference.

Henry Chauncey, *President*
Educational Testing Service

PREFACE

This volume provides a permanent record of the papers and discussion presented at the 1950 Invitational Conference on Testing Problems sponsored by the Educational Testing Service. The Conference was held at the Roosevelt Hotel, as usual, and the date was October 28, 1950. Attendance at the Conference continues to increase, and this year more than 250 makers, users, and students of measurement devices attended.

As has been usual in recent years, the program consisted of two morning sessions and an afternoon session. In planning these sessions, no attempt was made to organize them around any central "theme." Each session was an independent unit, developed in consultation with representatives of the Educational Testing Service as representing an area likely to interest members of the Conference.

The first session in the morning was planned as a follow-up on a session at the 1946 Conference. In the 1946 meeting, reports were presented describing a number of testing programs for the selection of professional personnel. This year representatives of four programs of professional testing, including some of the ones represented in the earlier session, were invited to report on the validation of their procedures and on their plans for future development of their program. This was thought of as a report of progress from which we would have a chance to see which procedures were working out well and which poorly. It was hoped that we might receive mutual stimulation from suggestions as to further types of selection procedures which should be developed.

For the second morning session we were fortunate to have Dr. L. L. Thurstone as a guest speaker. To some extent, the plans for Dr. Thurstone's participation "just growed," since a refusal to participate in another type of session was converted into an invitation to give a talk on creativity, a topic on which Dr. Thurstone indicated that he had been interested in organizing his ideas. I was indeed gratified when Dr. Thurstone accepted this second invitation.

The afternoon session was concerned with various aspects of the evaluation of educational achievement tests. We tried to think of the angles from which a test might be viewed and to get them represented in the discussion. Thus, Dr. Davis was asked to take the viewpoint of the test editor, Dr. Schwab that of the subject matter specialist, and Drs. Carroll and Gulliksen that of the statistician. Their approach was not quite what had been anticipated at certain points, but was in all cases provocative and stimulating.

After the initial arrangements had been made for Conference participants, the job of the Chairman was an easy and a pleasant one. The other arrangements were all most efficiently handled by the staff of the Educational Testing Service. The speakers were on hand at the appointed hours, with their materials organized and ready. The audience was large, responsive, and ready to participate in discussions within the limits of the available time. The hospitality provided by the Educational Testing Service was most agreeable. Many conferees were gracious in their comments on the program. I join with the many other participants in looking forward to future sessions.

ROBERT L. THORNDIKE, *Chairman*
1950 Conference

CONTENTS

PAGE

FOREWORD by Mr. Chauncey 5

PREFACE by Mr. Thorndike 7

PANEL I: "*Validation of professional aptitude batteries.*"

Tests for Accounting
Arthur E. Traxler and Robert Jacobs, *Educational Records Bureau* 13

Tests for Law
A. Pemberton Johnson, *Educational Testing Service* 30

Tests for Dentistry
Shailer Peterson, *American Dental Association* 35

Tests for Medicine
John M. Stalnaker, *Association of American Medical Colleges* and *Educational Testing Service* 46

Discussion 52

ADDRESS: "*Creative talent.*"
L. L. Thurstone, *University of Chicago* 55

PANEL II: "*Criteria for the evaluation of achievement tests.*"

From the Point of View of the Test Editor
Frederick B. Davis, *Hunter College* 73

From the Point of View of the Subject Matter Specialist
Joseph J. Schwab, *University of Chicago* 82

From the Point of View of Their Internal Statistics
John B. Carroll, *Harvard University* 95

From the Point of View of Their External Statistical Relationships
Harold Gulliksen, *Educational Testing Service* 100

Discussion 104

Appendix 113

PANEL I

Validation of Professional Aptitude Batteries

Validation of Professional Aptitude Batteries

ARTHUR E. TRAXLER AND ROBERT JACOBS

TESTS FOR ACCOUNTING

Except for occasional use of locally constructed tests in individual colleges, objective measurement in the field of public accounting is a very recent development. The first large-scale attempt at objective testing in accounting began in 1943 when the American Institute of Accountants appointed a Committee on Selection of Personnel to consider and to develop procedures for the improvement of the selection of personnel for public accounting. The committee arranged with Dr. Ben D. Wood to direct the project and designated the Educational Records Bureau as the operating organization.

After much discussion and preliminary exploration it was decided to experiment with three types of instruments: 1) an orientation test or a test of mental ability slanted toward the field of business; 2) tests of achievement at two levels—the end of the first year of the study of accounting in college and the end of the senior year of college or at the point of employment; and 3) a test of vocational interests.

The tests constructed or adapted for the project were described in a report at the 1946 Invitational Conference on Testing Problems, (12) and since that time reports on their use has appeared in various educational journals. (1, 3, 4, 5, 7, 13, 14, 15, 16, 17, 18) It will be necessary to review very briefly some of the information previously given in order to make the present report intelligible.

Nature of the Tests

The Orientation Test consists of three forms, each of which contains vocabulary and reading materials drawn from the field of business and arithmetic problems based on business situations. It requires fifty minutes of working time and yields a verbal score, a quantitative score, and a total score, which like the total score in the American Council Psychological Examination is simply the sum of the verbal and quantitative raw scores.

The Achievement Test, Level I, consists of eight sections designed to sample the work ordinarily covered in the first year of the study of accounting. It calls for a two-hour working period and provides one over-all score. There are now two forms of this test, and a third form is being prepared.

The Achievement Test, Level II, is intended to appraise the knowledge of accounting principles and procedures of graduating seniors or of applicants for accounting positions or of accounting employees. The first two forms constructed were designed for a four-hour working period. In response to requests for a shorter test, two two-hour forms have recently been made available. Each of these forms yields simply a total score. Part scores may be obtained, but these are not very reliable, and no norms have been established for them.

The Strong Vocational Interest Blank was chosen as the test of vocational interests. Special norms for public accountants were developed on twenty-seven occupational scales for that blank. A considerable amount of information concerning the use of the Strong blank in this project was given in the report to the 1946 Conference. The present report will deal, in the main, with the instruments constructed specifically for this project rather than with the Strong blank.

After a preliminary period of research on the new tests, two closely related programs were set up. These were known as the College Accounting Testing Program and the Professional Accounting Testing Program. The first College testing program was held in the fall of 1946, and testing in the colleges has been carried on each fall and spring since that time. The tests have been available to employers continuously since the spring of 1947. More than 130,000 accounting tests have been given in the colleges. In the spring of 1950, 204 colleges took part in the spring program and 175 firms participated in a special staff testing program. The results of all tests administered in the program are on file in the project office. Thus, an extensive accumulation of data is available for use in validity studies, provided suitable criteria can be found.

THE PROBLEM OF VALIDITY

We need to remind ourselves that the problem of validity is the crucial one in all test development. Beside it, the problem of reliability is relatively so simple and procedures are so well defined that there is no longer any excuse for test makers to release instruments without telling potential users very definitely, within the limits of defined populations and the application of particular techniques, what the reliabilities of their instruments are. For the tests under consideration it has been reported elsewhere that the medians of the Spearman-Brown reliability coefficients found with the scores of students at a given level of study are approximately as follows: Orientation Test, verbal score, .90; quantitative score, .80; total score, .91; Achievement Test, Level I, .94; Achievement Test, Level II, four-hour form, .97; Achievement Test, Level II, two-hour form, .85. These reliability coefficients are, in general, fairly satisfactory.

When one ventures into the field of validity, however, he enters a world that in its ultimate relations is not scientific but philosophical. As Rulon (9) pointed out at the 1948 Invitational Conference, we cannot expect to establish ultimate criteria of goodness. It is necessary to make value judgments of less than ultimate values if we are to avoid getting bogged down in a morass of metaphysical discussion. Nevertheless, for limited segments of human experience we can set up goals that will serve the purpose of ultimate criteria, assuming that our value judgments are sound.

Another difficulty with the study of validity is that the validity of a particular instrument is ever changing depending upon the setting in which it is applied. Validity is meaningful only in relation to the use to be made of the instrument.

A test may be considered from the standpoint of its validity with regard to the purposes the author had in mind when he constructed the test or from the standpoint of a user concerned with a particular purpose for which the instrument is to be used. In appraising the validity of the tests used in a program such as this one, it is necessary to consider validity in both these ways. In the first place, what is the validity of the tests for the measurement of those areas of experience they are designed to measure? The Orientation Test is intended to measure what might be called "business intelligence." To what extent does it actually measure mental ability as revealed in dealing with business materials? The Achievement Tests are designed to measure knowledge and understanding of the principles and procedures of accounting. To what extent do they measure this kind of knowledge and understanding?

This way of looking at validity is a common one, and the usual procedures may be used in studying this aspect of the validity of the instruments. As will be seen a little later, several familiar procedures of studying validity have been applied to the tests.

The second way of considering validity in a testing program of this kind is much more involved, for here we enter an area where the determination of criteria is a complex undertaking. For this battery of accounting tests, the "ultimate" criterion is prediction of success in accounting work. This is the end product in which both the colleges and the firms are primarily interested. They wish to select men of high caliber and to train them well so that their work as accountants will be superior in quality. So it is necessary to accept success on the job as a criterion of the validity of the tests. In the end, the program will succeed or fail largely on the basis of whether the users find that these tests help them choose better men for the accounting profession.

Even if a satisfactory criterion of success can be found, the interpretation of relationships between test results and the criterion is still a complicated one, for the accuracy of the judgment of the planners of the program in regard to what makes a successful accountant must be considered. It is reasonable to suppose that intelligence in business situations and knowledge and understanding of accounting are elements in the success of an individual accountant, but it is also reasonable to suppose that these are not the only elements in his success. Personal qualities, ability to deal with people, energy, writing ability, and other qualities are recognized as important for a public accountant. Correlations of hypothetical perfect measures of mental ability and achievement with a perfect criterion, then, could not be expected to be extremely high, for they, with their perfection, would measure only two of many factors which may condition success. But the correlations may reasonably be expected to be positive and fairly substantial. Correlations of less than perfect measures with an imperfect criterion may also be expected to be positive but of lower magnitude.

Earlier Data on the Validity of the Tests

An attempt has been made during the period of several years gradually to accumulate various kinds of evidence on the validity of the Orientation and Achievement Tests. It is not possible to summarize all this information within the time allowed for this paper. Some of the data obtained in the earlier stages of the project are shown in Tables I through IV.

It will be observed that Tables I and II show correlations between test

TABLE I

Correlations of Scores on Orientation Test, Form B, With Grades in First Year Accounting in College[1]

Institution	N	Grades	vs.	Verbal	Grades	vs.	Quan.	Grades	vs.	Total
		r		PE	r		PE	r		PE
A	189	.26	±	.046	.33	±	.044	.31	±	.044
B	363	.33	±	.032	.34	±	.031	.39	±	.030
C	61	.32	±	.077	.26	±	.081	.38	±	.074
D	272	.42	±	.034	.50	±	.031	.50	±	.031
E	166	.37	±	.045	.37	±	.045	.43	±	.043
F	258	.24	±	.040	.49	±	.032	.40	±	.035
G	69	.32	±	.073	.58	±	.054	.46	±	.064
H	181	.38	±	.043	.43	±	.041	.44	±	.040
I	129	.35	±	.052	.51	±	.044	.43	±	.048

[1] From Robert Jacobs, "Measurement and Guidance in the Field of Public Accounting," Doctor's Dissertation, Wayne University, 1949, as adapted from an unpublished report by Frances Triggs.

TABLE II

Correlations of Achievement Test Scores With Grades in Accounting Courses in Ten Colleges

Institution	Year of Study	Variables	N	r	PE
A	First	Level I-A Scores vs. Grades in Beginning Accounting	155	.60	± .035
B	First	Level I-A Scores vs. Grades in Beginning Accounting	359	.41	± .030
C	First	Level I-A Scores vs. Grades in Beginning Accounting	62	.63	± .052
C	First	Level I-A Scores vs. Grades in Beginning Accounting	54	.55	± .064
D	First	Level I-A Scores vs. Grades in Beginning Accounting	255	.64	± .025
E	First	Level I-A Scores vs. Grades in Beginning Accounting	95	.44	± .056
F	First	Level I-A Scores vs. Grades in Beginning Accounting	69	.59	± .053
G	First	Level I-A Scores vs. Grades in Beginning Accounting	125	.76	± .026
H	Second	Level I-A Scores vs. Grades in Second Year Accounting	52	.58	± .062
D	Second	Level I-A Scores vs. Grades in Second Year Accounting	93	.58	± .046
D	Second	Level I-A Scores vs. Grades in Second Year Accounting	56	.50	± .068
I	Senior	Level II-A Scores vs. Grades in Advanced Accounting	56	.35	± .079
J	Senior	Level II-C Scores vs. Grades in Advanced Accounting	146	.58	± .04
J	Senior	Level II-D Scores vs. Grades in Advanced Accounting	108	.66	± .033

scores and grades in accounting courses in a considerable number of institutions. It is well known that grades are not a very satisfactory criterion against which to correlate scores on objective tests, but at the college level they are usually the best available criterion, and, notwithstanding their limitations, they have probably been more widely used than any other criterion. The correlations between the Orientation Test scores and grades are not high, but, with one exception, they are statistically significant. It has been shown in another report that these correlations are, on the average, considerably higher than those between beginning accounting grades and scores on the American Council Psychological Examination, the Otis Self-Administering Test of Mental Ability, and the California Test of Mental Maturity, Short Form.[1] It is interesting to see that although it is known that the reliability of the quantitative score on the Orientation Test is considerably lower than that of the verbal score, the quantitative score has the closer relationship with success in the study of accounting as indicated by the assigned grades.

As one would expect, the Achievement Test scores are considerably more closely related to grades in accounting courses than are the scores on the Orientation Test. The fourteen correlations shown in Table II range from .35 to .76 with a median of .58. In view of the fact that grades usually are not very reliable and that they are likely to be based on factors other than achievement alone, these correlations are not unfavorable to the Achievement Test.

Correlations between scores on the accounting tests and certain other tests are shown in Table III. Correlations were found for the Orientation Test and four other tests of mental

TABLE III

Correlations Between Scores on Accounting Tests and Scores on American Council Psychological Examination, California Test of Mental Maturity, Short Form, Ohio State Psychological Examination, Otis Self-Administering Test of Mental Ability, and Graduate Record Examination[1]

School	Test	N	Orientation Test					
			Verbal		Quan.		Total	
			r	PE	r	PE	r	PE
A	ACE-Q-score	299	.25	± .037	.49	± .030	.34	± .035
	L-score	299	.72	± .019	.33	± .035	.69	± .020
	Total	299	.61	± .025	.46	± .031	.63	± .024
B	ACE-Total	94	.65	± .040	.56	± .048	.68	± .037
C	California							
	NL	87	.32	± .065	.36	± .063	.36	± .063
	L	87	.68	± .039	.40	± .061	.65	± .042
	Total	87	.67	± .040	.48	± .056	.67	± .040
C	OSU Psych.							
	Total	37	.83	± .035	.45	± .088	.80	± .040
C	Otis SA							
	Total	88	.49	± .055	.53	± .052	.56	± .049
			Achievement Test, Level II					
D	GRE							
	Gen. Math.	121					.461	± .048
	Phy. Science	121					.313	± .055
	Biol. Sc.	121					.242	± .058
	Social St.	121					.387	± .052
	Literature	121					.358	± .054
	Arts	121					.224	± .058
	Ef. of Exp.	121					.437	± .050
	Vocabulary	121					.307	± .055
	GEI	121					.499	± .046

---

[1] op. cit. Table I.

ability—The American Council Psychological Examination; the California Test of Mental Maturity, Short Form; the Ohio State University Psychological Test; and the Otis Self-Administering Test of Mental Ability. These correlations indicate a fairly substantial degree of relationship between the Orientation Test and other mental ability tests. Five correlations between total score on the Orientation Test and total scores on the other tests range from .56 to .80 with a median of .67. The relative magnitude of the correlations of the verbal and quantitative scores with scores on the other tests is consistent with what one would expect, except in the case of the Otis test which has a slightly higher correlation with the quantitative than with the verbal score. This result seems somewhat surprising in view of the rather large amount of verbal material in the Otis test.

The rest of Table III shows correlations between the Achievement Test, Level II, Form A, and the Graduate Record Examination. These correlations are positive to a significant degree, but the fact that they are not high suggests that the Level II test is measuring something rather different from what is measured by the various tests in the Graduate Record battery and that it may be a desirable practice to use this test as a supplement to the Graduate Record Examination when testing college seniors or graduate students.

An important criterion of success in the field of public accounting is scores on examinations for the CPA certificate. This is an extensive and difficult series of examinations of the essay type. Most states employ the CPA examinations prepared periodically by a committee of the American Institute of Accountants, but a few states, including Pennsylvania, use their own CPA examinations. The accounting testing program has not been carried on long enough for scores to be obtained on CPA examinations for many of the students tested in college with the accounting tests, but a small amount of data on relationships between these two types of tests are available. These are shown in Table IV. The correlation

TABLE IV

CORRELATIONS BETWEEN SCORES ON ACCOUNTING TESTS AND RESULTS OF CPA EXAMINATIONS OF THE ESSAY TYPE

Test	Method of Correlation	N	Correlation
Achievement Test, Level II, Form C, (Experimental) vs. Pennsylvania State CPA Examination	Rank Order	34	.67
Orientation Test vs. Institute CAP Examinations (excluding Commercial Law)	Product-Moment	40	.41

found between Achievement Test, Level II, and the Pennsylvania CPA examinations is fairly high. The correlation of the Orientation Test with the CPA examinations of the Institute is significantly positive although not high. Although one would expect the Level II test to be more highly correlated with the CPA examinations than is the Orientation Test, since both the Level II test and the CPA examinations are designed to measure achievement, the fact that a considerably higher correlation was obtained is probably due in part to the way the tests were administered. The Level II test and the CPA examinations were administered at approximately the same time, whereas the Orientation Tests used in correlations with the Institute's examinations were given in colleges at various times some years before the CPA examinations were taken.

Correlations Between the Accounting Tests and Ratings of Success in Accounting Work

It is difficult to obtain a satisfactory criterion of success in any vocation. Salaries earned might be one criterion, but this kind of information is not readily obtained from employers, and, even if it were available, it would not be a very suitable criterion, for salaries are likely to depend partly upon length of service, and perhaps upon personal impressions, as well as on job performance. Level of employment is another possible criterion, but, again, the level at which the individual is employed depends in part upon his experience and the length of time he has been with a firm. Actual records of production are probably the most reliable and most objective criteria of success on the job, but these are limited to occupations where output or results of work performance can be measured quantitatively. Ratings of job performance are often used as measures of success. Ratings are frequently too subjective and too low in reliability to be of much value, but in spite of their faults they are usually the best criterion that can be found. Their dependability can be improved to some extent by clearly defining the qualities being rated and through careful instructions to the raters.

Preliminary studies.—Experimentation with ratings was begun early in the accounting testing program. In one such study, accountants at three levels of work—seniors, first assistants, and second assistants—in a large public accounting firm were placed by their supervisors in one of three categories; 1) technically qualified and of good promise, 2) technically qualified and of some promise, and 3) technically qualified and of little or no promise. An experimental edition of Achievement Test, Level II, Form A, was administered to these employees, and the percentile ratings of the men rated in the different categories were studied. The relationships are indicated in Table V. It was felt that the progression from high

TABLE V

MEANS AND RANGES OF PERCENTILE RATINGS MADE ON EXPERIMENTAL EDITION OF ACHIEVEMENT TEST, LEVEL II, BY VARIOUS GROUPS OF EMPLOYEES IN LARGE PUBLIC ACCOUNTING FIRM

Group	Characterization	N	Percentile Ratings	
			Range	Mean
Senior	Technically qualified and good promise	3	84-100	91
Senior	Technically qualified and some promise	7	52-100	86
Senior	Technically qualified little or no promise	16	12-96	59
First Assistant	Technically qualified and good promise	11	46-99	79
First Assistant	Technically qualified and some promise	41	18-99	66
First Assistant	Technically qualified little or no promise	40	2-97	49
Second Assistant	Technically qualified and good promise	23	4-95	49
Second Assistant	Technically qualified and some promise	53	1-94	37
Second Assistant	Technically qualified little or no promise	26	1-49	15

percentile ranks to low percentile ranks as one goes from the higher rating categories to the lower ones indicated that the test had considerable value for the selection of promising employees in that particular firm.

A second rating study was made in the spring of 1948 in connection with experimental testing of items for a two-hour Level II Achievement Test. One participating firm submitted a combined judgment rating of twenty-nine employees with respect to technical competence by placing them in groups corresponding roughly to a normal distribution: approximately 10 per cent in the category indicating greatest competence, 20 per cent in the above-average category, 40 per cent in the average or middle group, 20 per cent in the below average category, and 10 per cent in the least competent group. The correlation between this ranking and the ranking based on scores on the experimental test was .59. It will be observed that in this rating procedure an attempt was made to direct the attention of the rater to a single factor, *technical competence*. In follow-up interviews it developed that rater-judgment was affected considerably by other factors which the raters felt were so important in accounting work that they could not be ruled out of any judgment of success. These "important" factors are personal qualities, such as energy, cooperativeness, and ability to deal with clients, plus what the public accounting employer calls "auditing sense." The last quality appears to be a kind of intuitive skill which enables the auditor to sense that something is wrong even when the books balance and everything seems on the surface to be in order. These intangibles make it especially difficult to set up a satisfactory rating procedure for public accountants.

A later study.—In connection with a staff testing program in the spring of 1950 an opportunity was created for a more extensive study of the value of accounting tests for the prediction of ratings of performance on the job. A total of 971 employed accountants distributed among 175 firms were tested with Orientation Test, Form A, and Achievement Test, Level II, Form C. The majority of the employees also took the Strong Vocational Interest Blank. The cooperation of thirteen of these firms was obtained in rating their employees with a rating scale which combined the features of several forms actually in use in some of the firms. The scale covered eight aspects of job performance. Five rating categories were provided for each area. These areas were identified by descriptive phrases ranging from "satisfactory" to "unsatisfactory" performance. The rating scale is shown in Figure 1.

The rating scales were scored quantitatively by means of a key, or stencil, on which numerical values were assigned to various positions along the scale. The midpoints of these values were arranged so that each of the five categories from lowest to highest was assigned a value of 1, 2, 3, 4, or 5.

The number of employees tested and rated in the thirteen firms ranged from 9 to 33, with a median of 18. Correspondence and discussion with representatives of the firms indicated that there were differences in the procedures used in the rating of the employees. In some of the smaller firms the rating was done by a single individual while in at least one of the larger firms the ratings represented the pooled judgment of the supervisors who had conferred extensively before the ratings were made. Because of the differences in the size of the firms and in the procedures used in filling out the rating form, the data from each firm were treated separately.

All the firms were asked to do the rating before referring to the test scores. It is known that in eight of the firms the rating was done without reference to the test scores, for the ratings were received before the scores were reported. The reports on test results were sent to the other five firms before the ratings were received. Contamination of the criterion is, therefore, possible in the data used from these five firms, although so far as is known instructions were followed with regard to making the staff ratings before using the test scores.

For each of the thirteen firms participating in the study, correlations were computed between test scores and ratings on Item 3 (knowledge of

RATING SCALE FOR EMPLOYED ACCOUNTANTS

This rating form shows eight qualities on which to rate each employee. The first four relate to knowledge and ability; items 5-8 concern personality; the last item is an overall rating. Please rate each quality as independently as possible. It is suggested that each one in the group be rated on Item No. 1, then each rated on quality No. 2, and so forth. Rate each individual by making a check (√) at that point on each scale where in your judgment he stands. The average or typical employee should be rated at or near the center of the scale. *Please do not attempt to rate a staff member unless you have adequate knowledge on which to base your rating.*

Employee ....

Classification ....

Date of Rating ....

1. Quantity of work	Unsatisfactory amount	Rather slow	Completes regular work in reasonable time	More than average amount of work	Completes a very large amount of work
2. Quality of work	Unsatisfactory performance	Sometimes inaccurate	Average level of accuracy	Above average accuracy	Work is of exceptional quality
3. Knowledge of accounting	Does not understand his work	Fair understanding only	Displays satisfactory knowledge of field	Above average knowledge of accounting	Displays expert knowledge
4. Ability to learn	Unable to learn	Learns slowly	Shows average ability in learning new jobs	Above average ability in learning new jobs	Learns at exceptionally fast rate
5. Dependability and integrity	Needs watching	Sometimes unreliable if he thinks he can "get by"	Honest and dependable. Works with supervision	Very dependable—works with only occasional supervision	Highly dependable. No supervision required
6. Initiative and responsibility	Lazy, needs prodding	Indifferent	Does ordinary assignments of own accord	Industrious and energetic	Does more than is expected
7. Cooperation—ability to get along with others	Inclined to make trouble	Gives limited cooperation	Usually cooperates willingly	Above average cooperation—well liked	Cooperates cheerfully—inspires others
8. Overall value to the organization	Low	Below average	Will be an average employee	Above average—may be promotional material	Superior—definite promotional material

Comments:

Signature of Rater ....

FIGURE 1

accounting), Item 4, (ability to learn), a subtotal of Items 1 through 4, a total of Items 1 through 7, and Item 8 (overall value to the organization). The correlations are shown in Table VI. The medians and ranges of the correlations are summarized in Table VII.

For reasons mentioned earlier, very high correlations between test scores and ratings were not expected. It will be seen that all the medians of the correlations for the various items are positive. Most of them are rather low, although several are substantial. So far as the Orientation Test is concerned, the quantitative score exceeds the verbal score in correlation with the criterion and is at least as closely related to the criterion as is the total score.

The correlations between the Achievement Test results and the ratings are higher than those for the Orientation Test scores and ratings. In fact, it would seem that in view of the nature of the criterion the correlations for the Achievement Test are fully as high as could be expected.

The median correlations for the thirteen participating firms as given in the lower part of the table show clearly that there are very wide differences among the firms. Apparently the results of the test did not line up well at all with ratings in firms 5, 6, and 11, whereas in the case of firms 2, 12, and 9, there was good agreement. In fact, in the case of firm 2 the correlation between the test scores and ratings is higher than one would expect the correlations between two separate ratings to be.

A question may be asked whether the prediction of the criterion could be improved through a combination of test scores. In order to study this question the total on ratings 1 to 7 was used as the criterion along with five other variables. These variables were verbal score on the Orientation Test, quantitative score on the Orientation Test, total score on Achievement Test, Level II, Form C, and scores on the accountant, CPA, and psychologist scales on the Strong Vocational Interest Blank. The psychologist scale was used because public accountants typically score very low on this scale.

It should be pointed out that this kind of use of scores on the Strong blank is not the most satisfactory use nor is it the way in which the results of the Strong blank are customarily used in the accounting testing project. It is the pattern of scores on the various scales of the Strong blank that is, of course, more important than the magnitude of the scores on any one or two scales. But it was thought that the interest scores on the accountant, CPA, and psychologist scales might in themselves have some relationship to the prediction of job performance.

Intercorrelations among the six variables were computed for the employees of each of four firms, and the Wherry-Doolittle method was applied to the correlations in an attempt

TABLE VI

CORRELATIONS BETWEEN SCORES ON ORIENTATION AND ACHIEVEMENT TESTS WITH RATINGS OF EMPLOYEES IN THIRTEEN PUBLIC ACCOUNTING FIRMS

		Orientation Test, Form A															Achievement, Level II-C				
Firm Code	N	Verbal vs. 3	Verbal vs. 4	Verbal vs. 1-4	Verbal vs. 8	Verbal vs. 1-7	Quant. vs. 3	Quant. vs. 4	Quant. vs. 1-4	Quant. vs. 8	Quant. vs. 1-7	Total vs. 3	Total vs. 4	Total vs. 1-4	Total vs. 8	Total vs. 1-7	Total vs. 3	Total vs. 4	Total vs. 1-4	Total vs. 8	Total vs. 1-7
1	12	.637	.550	.650	.550	.643	—.065	.115	0	.115	.014	.534	.496	.547	.496	.532	.658	.501	.443	.501	.442
2	13	.591	.729	.700	.671	.690	.709	.851	.791	.709	.781	.645	.791	.749	.677	.739	.733	.799	.758	.828	.763
3	18	.528	.527	.509	.201	.462	.482	.427	.448	.069	.417	.647	.634	.597	.201	.531	.367	.365	.353	.191	.374
*4	21	.422	.385	.506	.560	.434	.597	.278	.538	.589	.568	.459	.320	.491	.583	.435	.594	.504	.619	.529	.635
5	28	—.045	.124	.108	.088	.106	.152	.302	.381	.366	.298	0	.189	.181	.166	.163	.252	.389	.472	.401	.358
*6	18	.219	—.021	.154	—.043	.101	.319	.174	.263	—.014	.229	.341	.099	.268	.061	.223	.631	.460	.641	.465	.608
*7	9	.403	.420	.255	.117	.165	.408	.353	.369	.144	.225	.453	.361	.257	.110	.148	.619	.658	.636	.489	.546
*8	33	.248	.257	.282	.372	.289	.132	.177	.204	.317	.241	.221	.246	.279	.370	.294	.164	.337	.217	.274	.267
9	15	.642	.481	.591	.375	.480	.656	.758	.696	.608	.573	.679	.583	.642	.453	.521	.257	.052	.123	—.111	—.036
10	17	.246	.312	.274	.164	.144	.405	.133	.309	.449	.340	.354	.404	.403	.333	.295	.578	.510	.649	.583	.640
11	20	.203	.324	.178	—.015	.101	.336	.379	.256	.034	.133	.196	.336	.181	.015	.066	.519	.653	.575	.474	.534
*12	19	.536	.606	.634	.727	.575	.499	.600	.607	.805	.569	.597	.696	.682	.791	.610	.476	.550	.559	.668	.560
13	16	.142	.020	.084	—.217	—.031	.452	.511	.487	.444	.418	.136	.058	.091	—.118	—.003	.751	.638	.520	.482	.451

* Firms having report of results before submitting ratings.

TABLE VII

MEDIANS AND RANGES OF CORRELATIONS BETWEEN ACCOUNTING TESTS AND RATINGS

Median Correlations and Ranges of Correlations by Items

Item	Median r	Range of r's
Orientation Verbal vs.		
Item No. 3	.403	—.045 - .642
Item No. 4	.385	—.021 - .729
Items Nos. 1-4	.282	.084 - .700
Item No. 8	.217	—.217 - .727
Items Nos. 1-7	.289	—.031 - .690
Orientation Quantitative vs.		
Item No. 3	.408	—.065 - .709
Item No. 4	.353	.115 - .851
Items Nos. 1-4	.381	0 - .791
Item No. 8	.366	—.014 - .805
Items Nos. 1-7	.340	.014 - .781
Orientation Total vs.		
Item No. 3	.453	0 - .679
Item No. 4	.361	—.058 - .791
Items Nos. 1-4	.403	.091 - .749
Item No. 8	.333	—.118 - .791
Items Nos. 1-7	.295	—.003 - .739
Achievement Test vs.		
Item No. 3	.578	.164 - .751
Item No. 4	.504	.052 - .799
Items Nos. 1-4	.559	.123 - .758
Item No. 8	.482	—.111 - .828
Items Nos. 1-7	.546	—.036 - .763

Median Correlations and Ranges of Correlations by Firms

Firm	Median r	Range of r's
1	.501	—.065 - .658
2	.736	.591 - .851
3	.437	.069 - .647
*4	.517	.278 - .635
5	.185	—.045 - .472
*6	.226	—.043 - .641
*7	.365	.110 - .658
*8	.262	.132 - .372
9	.547	—.111 - .696
10	.347	.133 - .649
11	.229	—.015 - .653
*12	.603	.476 - .791
13	.280	—.217 - .751

* Firms having report of results before submitting ratings.

to select the tests that would give the maximum shrunken multiple correlation with the criterion.

It would extend this paper beyond reasonable limits to present and discuss all these data. The intercorrelations among the six variables for the employees in one firm are shown in Table VIII. The application of the Wherry-Doolittle method to the data from this firm indicated that the correlation of the Achievement score and the Orientation verbal score with the criteria was slightly higher than the correlation of the Achievement Test along with the criterion (.664 as compared with .640), while the addition of another test to the multiple would add more chance error than validity to the battery.

No generalization could be drawn as a result of the study of the multiple correlations between the test scores and the criterion for the employees in these four firms. Because of the marked differences among the firms in correlation of the scores with the criterion, no definite combination of test scores could be recommended for the prediction of ratings of job performance. For example, in another of the four firms, the combination of CPA interest score, accounting interest score, and score on Achievement Test, Level II, yielded the best prediction. It seems likely that notwithstanding the uniformity of the instructions given to the raters in all the firms, the raters in the various firms differed markedly in the weights

TABLE VIII

INTERCORRELATIONS OF SIX VARIABLES FOR EMPLOYEES IN ONE FIRM N = 17

Test	Orientation V	Orientation Q	Level II-C	Acct.	Strong CPA	Psych.	Rating Scale Items 1-7
Orientation Verbal		.446 ±.135	.559 ±.116	−.411 ±.140	−.105 ±.167	.161 ±.164	.144 ±.165
Orientation Quant.			.629 ±.102	−.338 ±.149	−.034 ±.168	−.362 ±.146	.340 ±.149
Achievement Level II-C				−.228 ±.160	.096 ±.167	−.040 ±.168	.640 ±.099
Strong Accountant					.500 ±.126	−.300 ±.153	−.001 ±.169
Strong CPA						.162 ±.164	.044 ±.168
Strong Psychologist							.022 ±.168
Ratings 1-7							

they subjectively assigned to different factors in assigning ratings to their employees.

PROJECTED RESEARCH

Two serious limitations in the foregoing study are the small number of cases available in each firm and the fact that the raters were untrained so far as the use of this rating scale was concerned. A study using a larger sample and an improved rating procedure is needed. Such a study, using approximately 250 cases in one large accounting firm, is in an initial stage. It is hoped that this study and other research will yield additional information concerning the validity of the accounting tests.

The Committee on Selection of Personnel, under whose general direction the accounting testing program is being carried on, is research minded, and is encouraging research not only in the project office but also in the colleges and universities participating in the project. Several institutions have reported to the committee that they have undertaken studies of the validity of the tests with their own students. It may be anticipated that a large amount of additional information on the validity of the tests will gradually be accumulated as the testing services to colleges and accounting firms are carried forward.

REFERENCES

(1) Budd, Thomas A. "The Selection and Educational Training of Cost Accountants," *Accounting Review*, Vol. 24 (April, 1948). Pp. 183-92.

(2) Budd, Thomas A. "Observations Concerning the Accounting Orientation Test at the University of Pennsylvania," *Results of Achievement Tests and Orientation Tests Administered in Schools of Business of One Hundred and Fifty-nine Colleges and Universities*, pp. 22-24. The American Institute of Accountants, College Accounting Testing Program, Bulletin No. 7. New York: Committee on Selection of Personnel, July, 1949.

(3) Caffyn, Harold R. "Practical Experiences with Scientific Selection of Accounting Personnel," *Journal of Accountancy*, Vol. 88 (October, 1949). Pp. 320-25.

(4) Caffyn, Harold R., and Traxler, Arthur E. "American Experience in Personnel Testing for Accounting Work," *Accounting Research*, Vol. 1 (July, 1950). Pp. 373-84.

(5) Committee on Selection of Personnel, American Institute of Accountants. "A New Yardstick for Accounting Skills," *Journal of Accountancy*, Vol. 87 (December, 1948). Pp. 451-57.

(6) Jacobs, Robert. "Measurement and Guidance in the Field of Accounting," Doctor's Dissertation. Wayne University, Detroit, Michigan. 1949.

(7) Jacobs, Robert. "Measurement and Guidance in the Field of Public Accounting," *Accounting Review*, Vol. 25 (January, 1950). Pp. 27-34.

(8) Nissley, Warren W. "Selection of Accounting Personnel," *Papers Presented at the Fifty-seventh Annual Meeting of the American Institute of Accountants*. New York: American Institute of Accountants, 1944.

(9) Rulon, Philip J. "The Criterion," *Proceedings of the 1948 Invitational Conference on Testing Problems*. Princeton, New Jersey: Educational Testing Service, 1949. Pp. 32-34.

(10) Schmidt, Leo A. *Objective Examinations in Professional Accounting* (Preliminary Edition). The American Institute of Accountants, Professional ~~Accounting~~ Testing Program, Bulletin No. 2. New York: Committee on Selection of Personnel, June, 1947. Pp. 28.

(11) Schmidt, Leo A. *Objective Examinations in Professional Accounting: Samples and Comment.* The American Institute of Accountants, Professional Accounting Testing Program, Bulletin No. 2A. New York: Committee on Selection of Personnel, August, 1947. Pp. 36.

(12) Traxler, Arthur E. "Project in the Selection of Personnel for Public Accounting," *National Projects in Educational Measurement*, pp. 51-64. A Report of the 1946 Invitational Conference on Testing Problems. Washington, D. C.: American Council on Education, August, 1947.

(13) Traxler, Arthur E. "An Objective Testing Program in Accounting," *Collegiate News and Views*, Vol. 1 (October, 1947). Pp. 1-4.

(14) Traxler, Arthur E. "Objective Measurement in the Accounting Field—A Second Report," *Collegiate News and Views*, Vol. 2 (October, 1948). Pp. 4-7.

(15) Traxler, Arthur E. "The College Testing Program for Accounting Students," *New York Certified Public Accountant*, Vol. 19 (June, 1949). Pp. 351-63.

(16) Ward, Lewis B., and Hassler, Russell H. "A Critical Evaluation of the Institute Personnel and Selection Testing Program," *Journal of Accountancy*, Vol. 90 (August, 1950). Pp. 113-20.

(17) Wood, Ben D., and Traxler, Arthur E. "The Use of the Strong Vocational Interest Blank with Public Accountants," *Certified Public Accountant* (July, 1946). Pp. 4-5.

(18) Wood, Ben D., Traxler, Arthur E., and Nissley, Warren W. "College Accounting Testing Program," *The Accounting Review*, Vol. 23 (January, 1948). Pp. 63-83.

Validation of Professional Aptitude Batteries

A. PEMBERTON JOHNSON

TESTS FOR LAW

THE STUDY of law aptitude tests, or more specifically law school aptitude tests, was begun at least thirty years ago at Columbia University. The first nationwide administration of a law school aptitude test did not occur, however, until February of 1948 when the Law School Admission Test was given to 2,753 candidates. The widespread use of the nationally administered test, despite its youth, is shown by two facts: 1) it has been administered to over 22,000 law school applicants in 2½ years, and 2) in the spring of 1950 score reports were sent to 92 different law schools as compared with 114 approved law schools listed by the American Bar Association in the *1949 Review of Legal Education.*

In examining developments which have taken place in this professional aptitude area let us consider briefly: first, the early law school aptitude tests; second, the steps which led to the development of the Law School Admission Test; and third, the reliability and validity of the Law School Admission Test.

EARLY LAW SCHOOL APTITUDE TESTS

Dr. I. L. Kandel in his 1940 monograph, *Professional Aptitude Tests in Medicine, Law, and Engineering,* and Dr. Dewey Stuit and colleagues in their recent book, *Predicting Success in Professional Schools,* have provided excellent surveys of the work with early law school aptitude tests. As early as 1921 Harlan F. Stone, then Dean of the School of Law at Columbia University, initiated experiments with law school tests, using at that time a test of abstract reasoning and symbols. In 1925 Dean M. L. Ferson of the University of North Carolina, aided by Dr. G. D. Stoddard, now President of the University of Illinois, began experiments at their institutions and at seven other universities regarding the Ferson-Stoddard Law Aptitude Examination. Another early user of law aptitude tests was Yale University, where Dr. A. B. Crawford and his staff developed their very effective Legal Aptitude Test. Apparently, Yale and Columbia have used law aptitude tests for a longer continuous period than any other institutions in the country. In the late 1920's and early 1930's, a number of law schools made use of the Ferson-Stoddard tests and obtained some encouraging results. The findings indicated that the test used with previous college record, personal interviews, and other information regarding the candidates' character, etc., gave very useful, practical results. In 1937 the School of Jurisprudence of the University of California adopted a locally constructed legal aptitude test for admission, and in 1940 Tracy reported upon an effective legal aptitude test at the University of Michigan. Since 1943, the authors of the Iowa Legal Aptitude Test have reported in several articles and in the *Examiner's Manual* some appreciable validities for the 1941 and 1946 editions at four to six law schools.

Development of the Law School Admission Test

Early in 1947 representatives of several leading law schools approached officers of the College Entrance Examination Board regarding the development of a Law School Admission Test, which would be available nationally several times a year. Following preliminary discussions, the cooperation of twenty-two law schools was obtained in the development of this program, which has been conducted since January, 1948, by the Educational Testing Service. The twenty-two schools now represented on the Policy Committee are:

Boston University
Columbia University
Cornell University
Duke University
Harvard University
New York University
Northwestern University
Rutgers University
St. John's University
Stanford University
Syracuse University
University of Buffalo
University of California
(School of Jurisprudence)
University of Chicago
University of Denver
University of Michigan
University of Nebraska
University of North Carolina
University of Pennsylvania
University of Southern California
University of Virginia
Yale University

The Policy Committee decided that this test should be an aptitude (ability) test as free as possible of items requiring specific knowledge or detailed information, yet predictive of scholastic achievement in law school. The test design specified by the committee precluded the use of certain types of items which were used in earlier tests, and which gave a decided advantage to examinees who were already somewhat familiar with legal terminology.

With the support of the Policy Committee, the Law School Admission Test program has been set up from the beginning to provide for a continuing program of test development, pretesting, item-type validation, and test validation. Selected test materials were pretested in the fall of 1947 at several cooperating law schools and about half of the items were chosen on an a priori basis for inclusion in the February, 1948, final form. A considerable amount of pretesting and of experimental test validation has since been carried out for the development of final form test materials.

RELIABILITY AND VALIDITY OF THE 1948 LAW SCHOOL ADMISSION TEST

A statistical analysis of the 1948 test by Dr. Frances Swineford has shown an odd-even reliability coefficient of .97 for a spaced sample of 500 cases. During recent months a number of individual school validity studies have matured for examinees who took the Law School Admission Test at one of the regular administrations during the first eight months of 1948 and who were later admitted to one of the Policy Committee schools. These studies have been conducted by Dr. Marjorie Olsen and Dr. William B. Schrader of Educational Testing Service.

Scores on the Law School Admission Test are reported on a standard scale with a mean of 500, a standard deviation of 100, and a range of 200 to 800. It is important in considering validity coefficients for the 1948 Law School Admission Test to recognize that the test was used for admissions selection. This led to a marked restriction of range of Law School Admission Test scores for admitted students at certain schools. For example, the standard deviation of admittees to one school was 60. In nine schools the standard deviations were all in the 70's.

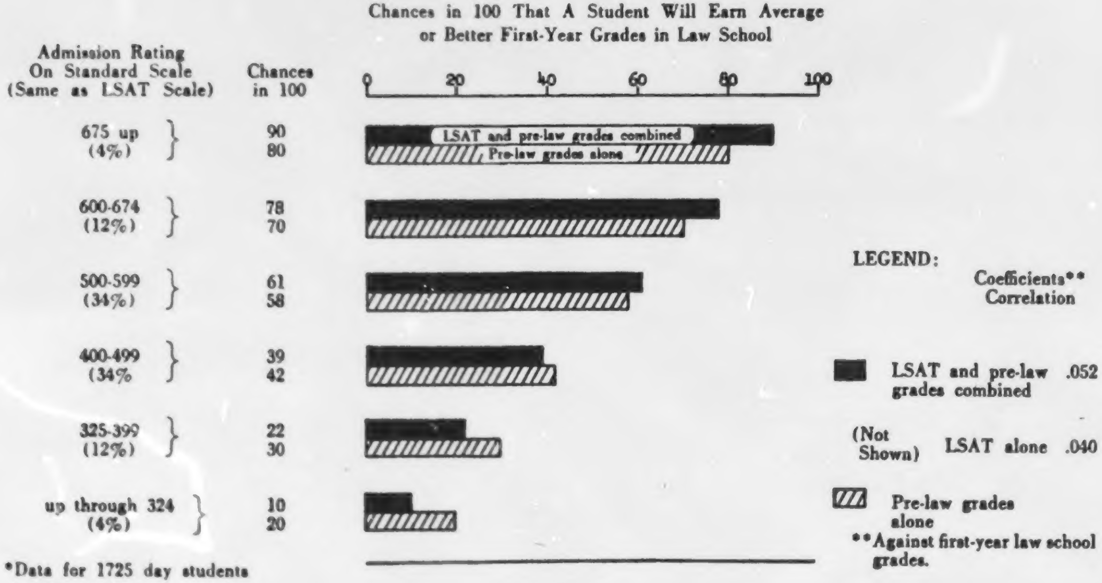
May I distribute to you at this point a figure and a table showing validity coefficients for the 1948 Law School Admission Test. Let us consider first the bar chart, Figure 1. It shows summarized data for twelve schools, the zero-order and multiple correlation coefficients being estimates of averages computed by Fisher's z-technique. If you look at the legend, you will note that the Law School Admission Test alone gives a slightly better prediction of first-year law school grades than do pre-law grades for this group of schools. The prediction of first-year law school grades by Law School Admission Test scores and pre-law grades combined is appreciably better than for either alone. The twelve law schools are:

Columbia University
Cornell University
Duke University
Harvard University
New York University
Northwestern University
St. John's University
Stanford University
University of California
  (School of Jurisprudence)
University of Michigan
University of Southern California
University of Virginia

Let us now consider the table. It shows the variation in individual school validity coefficients for Law School Admission Test scores combined with pre-law grades for pre-law grades alone and for Law School Admission Test scores alone. Note that the values in the column at the extreme right are adjusted by the usual correction for restriction of range, assuming a standard deviation of 100 for the unselected group. You will note particularly the stability of the multiple-R validity coefficients of

FIGURE 1

Effective Prediction of Scholastic Success in Law School
As Shown by the 1948 Law School Admission Test
and Pre-Law Grades Combined*



*Data for 1725 day students in 12 schools.

first-year law school grades from *school to school* when pre-law grades and Law School Admission Test scores are combined optimally at each institution.

Studies of the 1949 Law School Admission Test are in progress at several schools. At certain of these schools little, if any, change in their rather selective admissions practice has occurred between the two years. At other institutions, however, use of the test has: first, brought about greater confidence in its value; second, the requirement of higher scores for accepted applicants; and third, the resultant restriction in range of their scores. It has been reported to us informally that at one of the most selective schools the near equality in the high level of ability of the students has created a seriously disturbed grading situation. You will all appreciate how this may affect future validity studies at that institution.

A continuing program of individual school validity studies under a rotational system has been planned so as to cover, once in three years, each of the Policy Committee schools in which the N's are appreciable. Other validation studies and a considerable amount of pretesting are planned for this and the following years toward the goal of an even more effective Law School Admission Test program for the future.

1948 LAW SCHOOL ADMISSION TEST
INDIVIDUAL SCHOOL VALIDITY COEFFICIENTS
Criterion: Average First-Year Grades in Law School
Predictors: See Column Headings

Law School	(R) LSAT Scores Combined with Pre-Law Grades	(r) Pre-Law Grades	(r) LSAT Scores	(r_{adj})* LSAT Scores
Studies Based on More than 150 Day Students				
1	.55	.41	.44	.63
2	.50	.36	.41	.49
3	.50	.38	.40	.51
4	.48	.36	.37	.46
5	.53	.49	.26	.35
Studies Based on 100 to 150 Day Students				
6**	.68	.40	.65	.68
7	.51	.35	.35	.47
Studies Based on 40 to 100 Day Students				
8	.68	.50	.59	.70
9	.54	.34	.50	.60
10	.48	.42	.18	.26
11	.34	.19	.28	.37
12	.54	.36	.48	.55

The schools included above are: Columbia University, Cornell University, Duke University, Harvard University, New York University, Northwestern University, St. John's University, Stanford University, University of California, University of Michigan, University of Southern California, University of Virginia.

* Coefficients adjusted by the usual correction for restriction of range assuming a standard deviation of 100 on the Law School Admission Test for the unselected group.

** The test was administered to students in this group after admission to law school.

Validation of Professional Aptitude Batteries

SHAILER PETERSON
(DELIVERED BY CLARENCE M. PEEBLES)

TESTS FOR DENTISTRY

About eighteen years ago the American Association of Dental Schools indicated their interest in the use of aptitude tests for the selection of students by appointing a committee on aptitude tests to study this problem. Reports were issued from time to time indicating the studies that were being conducted at some of the dental schools in an effort to prepare testing instruments that could differentiate between a superior and an inferior dental student.

The Council on Dental Education of the American Dental Association is the accrediting agency of the dental profession, and one of its criteria in the valuation of dental schools is the admission standards. The Council became interested in the use of aptitude tests about ten years ago, and in 1944 named a committee to study this problem and included the members of the dental schools' committee on aptitude testing. In 1945 the Council employed the author of this paper to study the problem of aptitude testing as related to the field of dentistry, and he advised the Council and the schools in regard to methods that might be employed for improving the selection of students.

Up to this time there have been a number of small and independent studies conducted by individual teachers to investigate this problem of predicting dental school success. The two most significant studies that have been conducted were at the University of Iowa Dental School, and at the University of Minnesota. Smith, Greene and Freeman at the University of Iowa developed a short battery of tests, and these were given for a number of years at several cooperating dental schools. Douglas and McCullough at the University of Minnesota conducted a study utilizing testing instruments that were already in existence.

From a study of the literature, one could find evidence of nearly as many successes as failures in the field of predicting dental school success. Also, there was sufficient contradictory data available to cause one to question the test administration technics as well as the statistics that had been employed. The Council on Dental Education in cooperation with the American Association of Dental Schools decided to embark upon an experimental program of aptitude testing. All of the accredited dental schools in the United States volunteered to cooperate with the proposed program which was planned to extend over a five year period.

The first test battery administered in this experimental program was given in the summer and fall of 1946, and the fifth and last battery in the experimental program has just been administered in forty-one dental schools this fall of 1950. In 1948 the schools through their association agreed that the program was proving successful, and that the test data was useful to them. In view of this they requested the Council on Dental Education to continue the aptitude testing program and place it on a nationwide basis at the close of the experimental portion of the program. The nationwide testing program is being started in November, 1950, and all of the recognized dental schools are participating by requiring the dental aptitude tests of all students whom they may consider admitting to their 1951 freshman class.

The American Dental Association provided all the funds necessary for the conduct of the five years of the experimental program, and the Board of Trustees and the House of Delegates of the American Dental Association have now approved the operating budgets necessary for conducting the nationwide aptitude testing program, which has just been started. In the case of the nationwide program, fees are being charged the applicants, and these are expected to defray the cost of the new program.

The dental schools accept for their freshman classes each year approximately 3,250 students, and we find that there are approximately 11,000 applicants each year who file nearly 20,000 applications. The aptitude tests will be administered only to those students who have made application to an individual dental school, and who have been certified by that school as eligible for admission. Therefore, approximately six to seven thousand students of the eleven thousand who make application will be tested.

At the outset of the experimental aptitude testing program consideration was given to the aims and objectives of a program of this type. It was primarily intended to supplement the information already available by the admissions offices such as pre-professional records, university transcripts, interviews, and letters of recommendation. Secondarily, the objective of the program was to replace data gathered through interview, questionnaire, etc. Another primary objective of the program was to assist the schools in developing committees whose principal function would be to evaluate criteria for admission as well as criteria for evaluating the success of students and the success of teaching. Also, such committees would need to be concerned with curriculum development and the delineation of the aims and objectives of dental education.

All of the dental schools have committees and these have been an important and useful agency for the conduct of the aptitude testing program during its five experimental years. There is considerable evidence to prove that many of these same committees have become an important part of the educational programs in their own schools, and they have become a powerful influence in getting the various departments of the dental schools to evaluate their general course content in terms of their immediate objectives, as well as the general goals of dental education. In many institutions new and original research has been developed as a result of these committees, and indirectly as a result of the stimulus provided by the aptitude testing program.

The Council and its aptitude testing committee considered at the beginning of the experimental program the areas of testing which they felt needed to be considered in such a program. The abilities or skills represented are as follows:

1. Ability to read with understanding the type of material that he will be expected to read in dental school.
2. Ability to memorize verbal and visual material.
3. Knowledge of word meanings, both general and scientific.
4. Mental ability and ability to reason, as well as part scores on his linguistic and his quantitative ability.
5. Ability to visualize patterns and relationships without the necessity of preparing drawings of all these relationships. (Associated with this is the objective to measure the student's appreciation of artistic design).
6. General cultural background, by measuring the knowledge of current problems in a number of areas.
7. Ability to express himself orally and in writing.
8. Ability to use his hands and fingers skilfully and dexterously.
9. Interests.
10. Personality (or to secure some record of it).
11. Perseverance.
12. Ability to coordinate muscular movements.
13. Social instincts.

While all of these skills needed to be evaluated in such a program, there were definite limitations which had to be placed upon the selection of tests that were to be employed. In the first place, not more than one full day of testing could be required because of its interference with other portions of the program. The fact that the tests were to be given to students who had already been admitted to dental schools meant that the program would be administered after the regular school program had been started and, therefore, it would seem unwise to request the schools to devote more than one full day of such a testing program.

Another factor was the personnel of the individual committees. Many of the individuals on these committees had not had extensive experience in the administration of tests, and, therefore, an educational program was required to acquaint these committee members with some of the important procedures that must be followed in good test administration. This factor also dictated the principle that tests be employed which could be administered fairly easily and without requiring an extensive amount of preparation on the part of the test administrators. Because speed in itself was not one of the major objectives of the dental education program, and in order to keep the test administrative procedures at a minimum, it was decided to eliminate speed tests wherever possible and employ power tests.

It was decided that the battery should include tests that were already available and whose reliability and validity had been fairly well established. It was also decided that tests should be developed throughout the program designed to evaluate whatever specific skills and abilities required evaluation, and for which no adequate test instruments were yet available. It was also decided that some new tests should be developed in certain areas so as to point and direct attention more specifically to the field of dentistry than is customarily done in the available tests, in the field of reading, science achievement, etc.

It was also decided that the aims and objectives of the aptitude program be divided into two areas, namely, 1) the prediction of ability in the academic or theory field, and 2) the prediction of ability in the technical or clinical field. These two areas seemed to lend themselves to comparison and evaluation because many of the dental schools give courses that have a rather natural break between these two areas. In other words, grades are given to the students on their clinical work, and while this is integrated and therefore somewhat dependent upon the theory courses, nevertheless the evaluation of the student's work is primarily upon the basis of his technical ability. In order to predict this technical ability it was anticipated that tests of manual dexterity, ability to visualize patterns, and ability to follow directions and execute them manually were important, if not essential. In regard to the theory portion of the course in dental education, it was anticipated that tests of mental ability, tests in science achievement, and tests of reading skill would be important as methods of prediction.

The first battery of tests included a test requiring the student to carve in chalk some relatively simple geometric patterns. New forms of this particular test were prepared each year, and in most schools they have proven helpful in differentiating between the students who were able to demonstrate excellent technical or clinical ability and the students who were less able. Anothey type of test that has proven useful in predicting clinical or technical ability is the so-called "object visualization test." In some schools experimentation has been conducted with various types of the paper form board tests. As would be expected, there is a better correlation between these visualization tests and the manual dexterity or carving tests than there is between either of these tests and the mental ability tests. Therefore, the visualization and manual dexterity tests have proven to be useful instruments in predicting the success or failure of students in their technical laboratory work and in their clinical work.

The A.C.E. psychological test has been employed throughout the experimental program. It was selected not only because of the need for a mental ability test and the fact that it produced "quantitative" and "linguistic" scores, but also because it provided national norms based upon non-dental students. The committee had indicated its desire to have at least one test in the battery that would allow comparisons to be made from time to time between the students who are applying or who have been admitted to dental school and the current university and college populations.

Other types of tests which have been used and proven of value include specialized vocabulary tests, general vocabulary tests, tests in English composition, reading tests in the field of the natural sciences, science achievement tests in the fields of biology, chemistry, and physics.

In order to evaluate the effectiveness of the aptitude tests, or in other words, measure the validity of the tests as instruments of prediction, it was necessary to agree upon the criteria that would be used for measuring the validity of the tests. It was decided that it would not be practical to develop a series of achievement tests that could be used in all of the dental schools as a means of evaluating the achievement of the students. While the method of employing achievement tests as a criteria for evaluating the aptitude tests would have been desirable, the program for developing these and having them adopted by all the schools would have required considerable time. Therefore, it was decided to make use of the types of evaluation that were already available, namely the students' course grades.

These grades were found to be subject to the same limitations as those in any institution. Actually the shortcomings of the various grading systems served in many instances to be an advantage by creating an interest in educational research in many of the dental schools. A lack of consistency and reliability on the part of many of the grading systems is recognized by both the committee as well as by the instructors in the schools, as limiting factors in a study of this type. Some of the schools in which grading systems showed little or no discrimination between the best, the average, and the poor students, have changed their grading systems and methods. In these instances the improvements in grading have demonstrated improvements in the correlations between the aptitude test scores and the estimates of achievement.

Some differences in grading schemes and trends in marking can be explained on the basis of the philosophy of the instructor or the philosophy of the school, as well as on the basis of the actual students being tested. For example, the philosophy of some instructors seems to be to grade new students leniently and with very little variation for fear that poor grades might damage a potentially good student's later record. In some schools the dropouts and failures from year to year have a noticeable effect upon the over-all achievement of the class, and this is justifiable in terms of the resulting group being much more homogeneous. For this reason correlations between aptitude tests and grades are often much higher in evaluating freshman grades than when evaluating the upper-class students, in spite of the fact that the later grades are probably more reliable and frequently appear more valid for comparison.

The grades given to students in the technic courses as a general rule correlated high with both the carving dexterity test and the visualization test. Very few of the schools have clinical work offered during the first two years, but all have work that may be classified as pre-clinical which includes work in the technical laboratories and involves the carving of teeth, and the construction of sample dentures prepared in a rather elementary fashion. The zero order correlations between technic grades and the manual dexterity scores ranged from about +.1 to more than +.5, with the trend about +.4. In some instances these correlations were greater at the freshman and sophomore level than they were at the junior and senior level, and in about as many other instances the opposite was true. The differences seemed to depend upon the degree of homogeneity of the group. While different forms of the test were used each year there seemed to be relatively little difference between the various forms to predict success. The manual dexterity or carving test is the only test in the battery that cannot be graded by machine. The carvings are graded by a committee according to certain specific standards and in terms of relatively simple aims and objectives, namely the flatness of surfaces, the sharpness of the angles, and the symmetry of design that can be produced.

The zero order correlation between technic grades and the visualization test ranged from about .0 to more than +.5 with a general trend of about +.3. The intercorrelation between the carving test and the visualization test ranged in the different schools from +.2 to about +.6. The multiple correlations between technic grades and carving and visualization scores ranged from about +.3 to +.6.

It should also be observed that the correlations between either the carving test or the visualization test with academic grades were much lower than the correlations with technic grades. For example, in schools in which the correlations between visualization and carving with technic grades is more than +.4, the correlation with theory grades is less than +.2, indicating that the carving and visualization tests definitely have greater validity in predicting technic grades than they have in predicting grades in the theory or academic subjects.

Correlation coefficients between test scores and grades have shown in many institutions a general decrease since the beginning of the aptitude testing program in 1946, and in most instances this can be attributed to a greater homogeneity of the group, i.e. very little variation in test scores, in turn, has been brought about by greater care in the screening and selection of applicants.

Some of the highest zero order correlations between the test scores and the school grades are given in the following table.

TABLE I
Zero Order Correlations Between Test Scores and Dental School Grades

	Hi Highest correlation coefficients thus far computed with grades in:	
Names of tests	Theory	Technic
A.C.E. "Q"	.547	—
A.C.E. "L"	.623	—
A.C.E. "Total"	.651	—
G.E.D. Reading	.771	—
G.E.D. English Total	.648	—
Word Dexterity	.42	—
Michigan Vocabulary Total	.58	—
Survey of Natural Sciences—Biology	.407	—
Chemistry	.613	—
Physics	.509	—
Factual	.662	—
Application	.586	—
Total	.645	—
Carving	—	.489
Object visualization	—	.521
Predental G.P.A.	.54	.38

Multiple correlations as high as +.8 have been found between theory grades and the quantitative, reading, and science achievement. Some of the combinations of multiple correlations are indicated in Table II, as follows:

TABLE II

MULTIPLE CORRELATIONS OBTAINED IN SOME SCHOOLS BETWEEN DENTAL SCHOOL GRADES AND APTITUDE TEST SCORES

"Q"	"L"	"T"	Eng.	Read.	Bio.	Chem.	Phys.	Fact.	Applic.	Total Sci.	Carv.	Obj. Vis.	
		x		x									.729
x			x	x									.784
x	x		x	x									.786
x				x	x								.633
x				x		x							.790
x				x			x						.669
x				x				x					.771
x				x					x				.715
x				x						x			.764
x				x									.568
											x	x	.584

Regression equations have been developed from some of the available data but, of course, these equations are applicable only to those situations in which the data were collected. In view of the fact that there are some schools with less homgeneous groups of students, and with greater variation in their grading systems apparent, some of the potential possibilities of these test instruments can be observed and, therefore, sample regression equations can be prepared from these data. It is very likely that regression equations based on schools having these types of classes and these grading systems could provide prediction formulas whose basic value and importance might be used by all of the dental schools in considering the test scores of applicants now that the 1951 nationwide program is about to begin.

Most of the schools and the faculties in which the lower correlations are found between test scores and grades, realize the importance of the several factors that are causing these low correlations. The very fact that these high correlations can be demonstrated, points out the potential value of a test program of this type when it is possible to deal with a large number of applicants possessing a relatively wide range of abilities, rather than with a small number of students that have already been carefully screened before being admitted to the dental class.

All of the dental schools who have certified an individual applicant to take the aptitude test will receive a punch card record of his performance on the aptitude test. The holes in the punch card will provide a profile that is imprinted by the perforations on a printed graph. All test score codes are in single digits for punch card use and they employ a range of twelve points with the "4" position on the card as the "mean." Each point increment is in terms of one half a standard deviation. No total or over-all averages are provided on the card so as to encourage the admissions committees to evaluate the applicant in terms of the profile instead of on the basis of any single derived score that could be fictitious. The admissions committees have now become accustomed to this method of evaluation during the five years of the experimental aptitude testing program. In order to continue emphasizing the importance of the entire profile as opposed to evaluation of a single average, the use of regression type of equations or formulas has not been promoted. The admissions committees, upon receiving reports on the applicants, will be given suggestions as to the apparent relative importance of the various tests in predicting dental school success insofar as this has been observed in those schools in which these relationships could be determined.

It is recognized that a true aptitude test battery should probably include only tests that measure innate abilities. However, in batteries of tests that are used to select prospective students, it is often desirable to replace or to supplement academic records of required courses that have been taken, with tests that measure the student's comprehension of these courses. In other words, achievement tests often serve a useful purpose and, because they predict student success, may be included with justification in an aptitude test battery. It is for this reason that the science achievement tests were included in the dental aptitude battery along with such tests as the manual dexterity, mental ability, and visualization.

In regard to innate abilities, it was interesting that many of the dental educators felt that carving ability or manual dexterity might depend very greatly upon practice and training, and not be a so-called "innate" ability that could not be developed. They feared that the time at which this test was given might be a dependent factor. The carving tests which were administered in eight of the dental schools in 1946 to the admitted freshman, were readministered in 1948 to the same students who were then juniors. In the two and a half years of dental school the students had had considerable training and practice in manual manipulations and in carving. The correlations between the original test scores and the retest scores were about +.6 or +.7, and the mean score of the retest group was usually nearly the same as that on the original test. In most instances this difference was not significant at the 1% level. In general it would appear that students who were very poor in their carving ability at the outset remained reasonably poor in spite of training. While there was some improvement in many instances, even the students who had previously demonstrated average or above average ability, were not, after the two years of training, able to demonstrate top level performance. Therefore, it would appear that perhaps there is some critical score representing a low level of ability, and if students receive scores below this critical level they are very apt not to be able to benefit from training and experience to become sufficiently competent for their dental school work. Students with such poor scores should not be encouraged to enter a career of dentistry. The level of this critical score appears to be about two and one-half standard deviations below the mean of the entire freshman group. High scores which the student may receive in his other tests apparently will not compensate for low ability in manual dexterity.

A score two and one-half standard deviations below the mean will appear so low to most people that its significance as a so-called "critical score" may be misunderstood unless further explanation is given. It must be kept in mind that this score is two and one-half standard deviations below the mean established by students who had already been admitted to dental school after being screened and selected on the basis of a number of criteria. Therefore this score becomes most useful and effective when it is considered as a "critical score" to be used when the tests are given to applicants seeking admission. In other words, while this test apparently measures innate ability to a rather large degree, students whose scores are higher than the "critical score" mentioned may actually be able to succeed in dental school although those with the lower scores obviously will have more difficulty. Those with scores below the so-called "critical score" perform so poorly that apparently it is impossible to improve their carving ability or manual dexterity sufficiently through training and practice to enable them to succeed in dental school. It should also be borne in mind that this may be considered a "critical score" in this one single criteria, whereas applicants may also be eliminated from consideration by exceedingly low scores in one or more of some of the other tests. When one considers again that the scores under consideration are those of students who have been admitted and that only about four per cent fail or withdraw during the first year it is quite obvious that the "critical score" based on such a group of students must be exceedingly low.

Studies have been conducted on the students who have failed, withdrawn, or been dropped from the dental school curriculum, and in most cases it has been possible to obtain an estimate from the administration as to the reason for a student being dropped so as to differentiate between scholarship failures and the withdrawals or drops that cannot be attributed to lack of scholastic ability. In a high proportion of instances it has been demonstrated that most of the students whose failure or withdrawal has apparently been due to lack of demonstrated scholastic ability, would not have been admitted had aptitude test scores been available at the time of selection.

As would be expected, there are cases in which students who obtain a low aptitude test score demonstrate a great deal of ability in their dental school program. In many of these instances it has been demonstrated that the students, when they took the examination which was given after they had been admitted to the dental school, did not attempt diligently to demonstrate their real competence and, therefore, their test scores are gross underestimates of their true potential ability. Motivation is recognized as a factor in all testing programs, and the fact that the tests were all given after the students had been admitted, and in some situations given by inexperienced personnel, would necessarily affect the attitude of the students and, therefore, make the grades of some of them inadequate in predicting future success.

In spite of the limitations placed on the dental aptitude testing program during its five-year experimental period, it is gratifying to report that much positive evidence has been secured. It is also gratifying to know that the program has had value beyond that of furnishing research data in the field of educational measurement by providing a stimulus and promoting educational research in many of the dental schools in the United States. The program has also been of interest in the state dental licensing boards, and during the last two years, programs for studying improvements in testing technics have been started and conducted by dental school teachers and by members of the state dental licensing boards.

Some reports on the findings in individual schools are now being prepared for publication. A monograph describing the entire program and printing complete data for the various schools will be published by this author in about two years.

Validation of Professional Aptitude Batteries

JOHN M. STALNAKER

TESTS FOR MEDICINE

I HAVE been asked to describe the work of selection of students for admission to medical school and to speculate on the needed developments. After one year of working on these problems, I am unable to suggest an easy answer to how the students should be selected or even to outline the research which appears to offer the most promise of help in solving the selection problem. I can report some of the conditions which seem important, some of the efforts being made and some of the research and development under consideration.

The problem which confronts each of the 79 medical schools in this country, and which is being attacked with a sincerity and effort seldom equaled, is that of selecting from a large applicant group the ones who are to be admitted to the study of medicine. This situation has not always existed. There was a time, well within the memory of some of our elders, when medicine was not the highly respected profession it is today. In the early 1900's only a high school education was required for admission to medical school and even that requirement was not strictly enforced. Medical schools were numerous. No laboratories or equipment were required. Training medical students was financially profitable even though the fees charged were small by today's standards. Students were sought, not turned away. There was no selective admission and failures were few in the course which lasted less than twelve months. Medicine and medical education have come a long way since 1900.

For the class entering medical school this year — 1950-51 — there were over 22,000 individuals who made application. Indeed they made almost 82,000 applications. Seven hardy young men each applied to over 35 medical schools. While more applicants were admitted — 7253 — than at any time since medical schools were reformed in 1910, some 15,000 aspirants were turned away.

Two of our schools each received over 3000 applications for the 125 places in their freshman classes. Think of it: 24 applicants for each place. At the other extreme, there were only 100 applicants for one school or about 3 for each vacancy.

Considering such figures as these, one can appreciate why the public concludes that admission to medical school is restricted to the brilliant applicants of straight A record. Such a conclusion is false, however, for a great many poorly qualified students apply. While high scholastic records for the accepted group are the rule, it is estimated that approximately 10 per cent of the group admitted this year had undergraduate records averaging below a B minus. Some 5 per cent of the group admitted received a score on the Medical College Admission Test — verbal section — the national percentile rank for which can be expressed with a single digit. The pool of applicants is not all cream.

How then are medical students selected? They are selected with a diligence requiring unbelievable amounts of time. Working with sincerity and utilizing the best techniques suitable to their situation, most of the admission groups are selecting a freshman class which appears to be superior relative to the applicants available to that school.

Like all other types of educational institutions, medical schools differ greatly one from another. There is good reason for this variation. The mean score on the verbal factor on the MCAT for the admitted group to one medical school is 2.3 sigma above the mean for another school's admitted group.

Selection is usually based upon the following types of evidence, each of which will be commented on briefly:

1. Biographical information as obtained from an application blank. Probably two items are of chief importance here: sex and home residence. Although only three medical schools exclude women (and one excludes men) many schools limit the number of women accepted. One school — Meharry — restricts its enrollment to Negroes. Some 20 odd southern schools exclude Negroes. About 2.6 per cent of students enrolled in medical schools in 1949-50 were Negroes. They were spread among 42 schools.

Last year 19 medical schools restricted admission completely or with two exceptions at most to residents of the state, thus lowering in some cases materially the quality of the accepted group.

Other items of biographical information may be considered but few if any schools make a systematic effort to evaluate biographical data. Sons of physicians and of alumni may be given some preference in a few schools but even in these cases the weight is less than popularly believed.

2. Academic record is given the principal weight by almost all schools. The courses taken, the grades received, and the standards of the undergraduate college are all considered.

3. Recommendations are considered seriously by some schools and promptly filed in the wastebasket by others. One school rates all letters on a numerical scale roughly proportional to the strength and number of the adjectives used. Other schools consider only letters supplying sup-

porting evidence for their statements or letters from sources of established reliability. There is a tendency to favor recommendations from committees of undergraduate advisers and professors.

4. Test scores while treated with a wide range of respect are coming to be used more and more sensibly. Both their values and their limitations are being recognized. One or two schools, however, say they pay no attention to the scores. Many schools place almost as much weight on the test scores as on the college transcript. The tendency to use test scores in place of thinking is being combatted by offering suggestions concerning the use of test scores as an aid in the interpretation of a total record. While the test is believed to have some value in screening out the less suitable, high scores alone should not be considered as a designation of the proper type to be admitted. The present test consists of four sections: verbal, quantitative, understanding modern society, and science.

5. The interview is widely used. Able senior professors, including those from the clinical departments, are devoting hours of costly time to this work. In some cases the interview is used as a device for integrating all available information about an applicant. It provides an opportunity for the interviewer to probe into areas not adequately covered or to check any apparent inconsistencies. Following the interview an over-all rating covering all evidence is made. A great many variations are found in how the interview is handled and the aims of the interview. Some interesting data on the stability of one type of interview is now being obtained at Iowa. Another type of study is under way at Utah. The interview, some of us think, unfortunately, is also being used in some instances to assess the psychopathology of the applicant, his basic motivation both its nature and its strength, as well as such worthy if hypothetical traits as broadmindedness, ability to work cooperatively with others, moral and ethical standards, ambition, and love of mankind. Trick interviews and group interviews and many other types are being used. The degree of success of these interviews awaits careful study.

The interview is usually aimed at assessing important characteristics not adequately covered elsewhere. It must be improved.

6. A physical examination has recently been added by some schools.

There is no uniformity in the way in which information obtained from these six sources is combined to produce an answer of accepted or not accepted. In at least one school mathematical weights are assigned to each of the several items and a composite rating obtained, available for legislative inspection. In general, however, each case reaching the final selection committee is accorded careful review and what may be called a clinical type of judgment made. This latter type of approach, I regret to say, seems to me the appropriate one at this stage of development for the admission groups now working on the problem.

So much for what is done. What should the procedure be?

According to my teachers, ideally one should consider the objectives of the training, devise instruments to measure the necessary background conditions which will facilitate the attainment of these objectives, measure the degree to which the applicants possess this ability and aptitude and then select those possessing the greatest amount. Along this logical, appealing, and in the long run essential, approach, lie many pitfalls.

While the M.D. degree is necessary for success in the fields we are considering, it is not sufficient. Grades alone are not enough. The admission officers are not focusing their attention on traits necessary for success in first year of the study of medicine or indeed in the total medical school program. They are attempting in many cases to select the man believed to possess in at least embryo form those characteristics and attitudes deemed desirable in the professional life of the M.D. Yet if we look to the products several years out of medical school, what do we find? They follow many and very different lines of work common only in that success in any of them is difficult to measure. Medical schools are training men who will become: general *practitioners* (family physicians), *specialists* (ranging from otolaryngology to proctology, from preventive medicine to plastic surgery), *research workers* on almost all phases of medical research, teachers and *medical educators, administrators* (within the medical fraternity there continues a belief that only an M.D. can handle properly administrative work connected with medicine including hospital management) and we must add medical *politicians*, a growing vocal group which may include an occasional statesman. I refer to men doing work in the county, state and national organized medical groups. These groups—practitioner, specialist, educator, administrator, politician, research worker—I believe are very different breeds and probably fundamentally different in basic characteristics. If the admission committee is trying to pick not the best of each breed but the best of the show it has an impossible task. There is room for many types in medicine.

Even if we assume we are seeking a single type, how can we define success in the practices of any aspect of medicine — the psychoanalyst, the chest surgeon, the hospital administrator, the teacher, the research worker? Is it the amount of money earned, the fame gained, or the judgment of one's peers? While these questions are not beyond answering, no simple general answers have been found.

The usual procedure, therefore, is to resort to grades in courses in medical school, on the assumption that high grades predict success in anything and everything.

Two points may be made about grades in medical schools. First, the grades given in professional schools may have a special meaning. In the two preclinical years where basic science courses are usually taken, medical schools have one teacher for each 4 to 5 students. In the two clinical years, one teacher is used for each 1 to 2 students. Some schools have more full time teachers (or their equivalent) in the clinical years than they have students. Many—most—of these clinical teachers are part time and many are voluntary, i.e. unpaid. Grades given under these conditions may have special meaning.

The second point is that every effort is now being made to retain students and pass them from year to year. We need more doctors; we need more good doctors.

Before grades are given too great weight, they should be studied further.

While I should be unwilling to discourage anyone from correlating any two variables, I am neither impressed nor concerned when a low correlation is found between scores on a test in understanding modern society and grades in laboratory work in gross anatomy. I continue to favor selecting the men for the study of medicine who have some awareness of social sciences. Only this morning I noted in the *New York Times* that Dr. Paul R. Hawley, director of the American Colllege of Surgeons, was quoted as saying "the difficulties in which the medical profession now finds itself are due, in large part, to the failure of doctors to realize that a social revolution has been in progress in this country for three decades, and that it has undergone considerable acceleration in the past seventeen years. . . . profound social changes have already occurred which neither we nor anyone else can reverse. . . . [Our only hope lies in an] honest effort to adjust free medical practice to these changes." Correlations or no correlations, we want men entering a medical career to have some training in the social sciences.

If the average grades in first-year medical school do not correlate highly with a score which may crudely be representative of intelligence, I shall not conclude that a stupid M.D. is as good as a bright one as far as diagnosis of disease in my personal family is concerned.

The tests we use now can be said to have some intrinsic validity. Within certain limits they can be used with confidence, not as a sole criterion, but as one item of evidence. Frankly, I am impatient with those who say the test is experimental. Until it is "proved" they rely on letters of recommendation, the interview, and grades from an unknown college.

Time does not permit the development of these fuzzy ideas. In concluding may I say we are working in the following areas:

1. A study of all students who have dropped out of any medical school during the past five years.

2. Tests of interests suitable for pressure situations where the applicant seeks to convince, not to reveal.

3. The development of device to detect serious emotional instability, again suitable for pressure conditions.

4. A study of the interview and what it can yield is under study at several locations.

5. A study of the nature of the students entering two of our medical schools, utilizing psychiatric techniques.

6. Routine correlational studies.

How can we select the best for admission to our medical schools? The problems are many. The task is complex and difficult. On the brighter side, the quality of medicine practiced in the United States is higher than that of any other country. There is some reason to believe that the selection of students for medicine is improving. After going over the credentials of the group admitted to one of our medical schools recently and then meeting the group, I came away with the impression that whatever the procedures being used at that particular school, the results looked very favorable. Would that we were doing as well in selecting graduate students in education and psychology!

DISCUSSION

PARTICIPANTS:

Robert L. Thorndike, Hugh M. Davison, Arthur E. Traxler, John C. Flanagan, John M. Stalnaker, Clarence M. Peebles, Guy E. Buckingham.

Chairman Thorndike: This session this morning is listed as a symposium, and a symposium implies that the audience is invited to participate. You have heard the contributions of our four representatives. As is typical of symposia, those presentations have taken a little more time than they were originally scheduled to take, but now we will at least give a few minutes' opportunity for those of you who would like to address a question to one of the members of the panel, or make a brief comment on some aspect of what they have said.

Dr. Davison: I think there needs to be some generalization related to this whole problem of fitting people for vocations. I wonder if some member of the panel could not give that to us. I noted several trends within the discussion.

Chairman Thorndike: Does any member of the panel wish to generalize the trends which are emerging from their respective remarks?

Dr. Traxler: Mr. Chairman, I thought that was the Chairman's problem.

Chairman Thorndike: One theme that seems to come up again and again is that grades are horrible criteria. Another theme that seems to come up again and again is the effect of restriction of range. Beyond those, there is a realization that life is more complicated than you think.

I don't know that that quite adequately summarizes the discussion.

Dr. Flanagan: I had one question to ask John Stalnaker. I keep reading and working to try to find a way for selecting people. All of my experience and reading indicate that the clinical method of making selection decisions is to be avoided. The reported studies indicate that almost any objective procedure, like grades or tests, will give a better prediction than you obtain if you let somebody review and interpret the data and weight the various results in accordance with their own thinking. I just wonder if there is some new procedure or some new data which would indicate that I should reverse my ideas of trying to get systematic judgments made and put into quantitative form and of getting people in rank order and selecting from the top rather than using "clinical procedures."

Mr. Stalnaker: I think you know I wouldn't ask you to change your procedures.

All I mean to say is that a job is being done; it must be done; a class has to be selected. At the present time I would like to see the multiple regression formula approach held in the background for just a little bit. At one institution they take the grade point average, no matter where the person went to school. Consider the illustration of two boys interested in the study of medicine; one took a course in appreciation of music, and so he is able to go to medical school; the other one took a course in physical chemistry, so he is able to appreciate music now.

I think it is not sensible to take a grade point average of a miscellaneous type of institution and weight that mathematically by multiplying it by seven and saying, "That is the factor." The same institution takes the letters of recommendation. They are reviewed by three people and they are given a rating as extremely enthusiastic, which is 40 points, enthusiastic is 20 points, and the others are 10 points. I think that is nonsense. I don't like it. I would like to have the quantitative approach made, obviously, but I want the current group in medical admissions to do a little more thinking about their procedures, and until we get the thinking done, I am inclined to accept what I called here clinical judgment. In the long run, I hope to achieve what you are achieving now.

Dr. Peebles: I should like to say "Amen" to what Mr. Stalnaker has just said. Furthermore, under the conditions that have persisted the last few years, the incidence of failure which I would call due to lack of ability and aptitude is very low. In our institution I know it is less than 2 per cent. We have more students falling by the wayside now because of emotional disorders and other difficulties, such as money, marital problems, and so forth.

Dr. Buckingham: It seems to me that we have shown, or some people have shown, that there are certain limitations on validities for predicting and in working on the other end of the process to refine the measuring stick of the professor. I would like to ask Mr. Stalnaker if his work connotes dealing with the methodology of teaching in medical school.

Mr. Stalnaker: I was supposed to stick to fifteen minutes in what I had to say, and I didn't want to reform medical education in that time, but I would say that the whole problem cannot be properly approached unless you do consider what happens to the student in medical school. I think you have to consider both the background of the student, his abilities at the time you are admitting him to school, and what process he is going to be subjected to. So I would answer yes to what you have implied.

DR. FLANAGAN: I would like to make one more comment. We have been advising the Dean of the Pittsburgh Dental School on how to select his students. He trustingly has accepted our judgment as to how to do it. We have studied the requirements of dentistry as reported from men in the profession, as contained in incidents reported from training schools, and so forth, and have assigned weights, of course not multiple correlation weights, because we haven't adequate criteria and lack five or ten years of experience. We have assigned weights based on the requirements as indicated by these reports to grades, to pencil and paper test scores, and to scores on a chalk-carving test. We have tried to take into account other things by a half-hour interview with a member of the committee, with the usual lack of very conspicuous success; correlations of the interview scores seemed to be positive with later ratings but are very small.

However, we seem to have all the information that we can get, including rating of recommendations. We assign numerical scores, and I am happy to relate that the admissions committee have numbered them 1 to 100 and filled the quota by starting from the top and going down.

ADDRESS

L. L. THURSTONE

CREATIVE TALENT

ONE OF THE most important educational problems is to encourage creative talent, to learn how to find it, and how to train it. We need creative talent in all fields of life. Creative talent is in demand in science, the fine arts, industry, and in the professions. In comparison with the social importance of this kind of talent, we have not done much to learn about it.

At the very beginning of this paper we must admit that little is known about creative talent. We have in mind here the intellectual and temperamental traits that characterize people who are able to formulate problems and to find original solutions. Because of the importance of creative talent, it seems worth while to study it experimentally. We shall consider a number of hypotheses about the conceivable nature of creative talent and the types of observation and experiment by which it might be investigated.

Let us first try to delimit the problem as far as we can in the present state of knowledge. Perhaps the first thought that comes to mind is to investigate anecdotal material concerning the work of geniuses. If we should focus attention on such studies, we should also select some living geniuses and study them at work. To limit investigation to the study of geniuses is probably not a profitable way to proceed.

Let us begin with the working hypothesis that creative talent is qualitatively the same at all levels: in the trades and in the professions, as well as in the rare and extreme forms that we call genius. Most men in the professions and in business do creative thinking. Of course, some do more than others. We assume that the thinking which leads to the solution of problems at the professional level is qualitatively the same as that which characterizes the work of geniuses. The difference may be one of power. If this working hypothesis is sound, then we can learn much about creative thinking by studying the working habits of professional men and the conditions that seem to favor the successful solution of problems. Such study might be cooperative. We might study our work habits in relation to successful problem solving. A few hypotheses will be mentioned that we might keep in mind in watching ourselves at work. Such inquiries can lead to clues that may eventually be useful for understanding creative work, whatever its nature may turn out to be.

Before eliminating the upper extreme of genius from our inquiry, let us correct a common misunderstanding about creative talent. It is rather common in the field of mental measurement to discuss the distribution of intelligence in terms of a normal curve. At one end of this curve we indicate the low intelligence of those unfortunates who do not have enough mental endowment to get along in competitive society. They are the idiots and the imbeciles. Then, at the upper end of the distribution it is rather common to indicate the range of genius. This is a serious error. If genius represents extremely high gifts for creative thinking, then it is not synonymous with intelligence. To be extremely intelligent is not the same as to be gifted in creative work. This may be taken as a hypothesis.

It is a common observation in the universities that those students who have high intelligence, judged by available criteria, are not necessarily the ones who produce the most original ideas. All of us probably know a few men who are both creative and highly intelligent, but this combination is not the rule. One does not expect to find creative talent in a stupid person, but there may even be some doubt about this statement because creative talent in one field can be associated with general mental endowment that is mediocre. There is undoubtedly a positive correlation between creative talent and intelligence, so that geniuses are usually in the upper half of the general intelligence distribution.

The confusion between intelligence and creative talent is common. For example, Quiz Kids are often referred to as geniuses. They would undoubtedly score high in memory functions, including incidental memory and rote memory. But it is doubtful whether they are also fluent in producing original ideas. If this really were the case, then the problem of selecting creative talent would be simple. Other examples are the students who complete their college degrees at an early age. Experience does not indicate that they later produce the ideas that advance our civilization. On the other hand we should expect to find an occasional genius among them.

Those who supervise Doctor's and Master's dissertations in the universities often have occasion to speculate about this problem. A Doctor's dissertation is, by definition, "a contribution to knowledge." As instructors we work hard to make an occasional contribution to knowledge, but it is difficult to insist that the inexperienced student do likewise. There is good reason why dissertations are often referred to as "pieces" of research—often trivial pieces. Although a tremendous volume of dissertation research is being poured out from the universities every year, it does not offer any solution to the problem of defining creative talent.

It will probably be agreed among experienced teachers that it is not the student with the highest scholarship who produces the most interesting and original ideas in the laboratory. Students who have creative ideas are sometimes erratic and undependable. They might argue that scholarship represents a certain degree of intellectual docility. Although there seems to be some conflict between scholarship and creative talent, they are probably positively correlated. Several primary factors of memory have been indicated in factorial studies. Incidental memory seems to be independent of the ability to memorize paired associates, and these seem to be different from visual memory as primary factors. Men who excel in producing original ideas should be studied as to the several memory factors.

It might seem plausible to formulate the hypothesis that people who are creative tend to be generous towards ideas. Perhaps they should be characterized as open-minded. Although this hypothesis has some face validity, it will probably be found that creative people do not excel in open-mindedness.

By way of further delimiting the problem we might consider the hypothesis that scientific talent and creative talent are not the same. It seems possible for a man to have good scientific talent without being creative. By scientific talent we mean here the ability to handle competently the concepts and methods of a science or profession without the ability to produce new ideas that are commensurate with mastery of subject matter.

Let us turn next to a tentative formulation of differentiating criteria for the creative act. The solution to a problem is usually characterized by a moment of insight in which the significant question is asked and perhaps solved in the same moment. The moment of insight is the critical moment. The thinking that precedes the moment of insight is different from the thinking that follows that moment. We might define the moment of insight as the main characteristic of work that is called creative. It does not necessarily follow that a problem which is so solved is important. A problem that is solved by the creative act of one man may be solved quite casually and routinely by another man with different experience and training. It is the nature of the psychological processes that should determine whether we call the act creative. A creative act may be dead wrong. That is not known until the new idea has been tested.

It is in the nature of creative work that it has some novelty or invention. Whether society considers the idea to be novel has nothing to do with the case. The act is creative if the thinker reaches the solution in a sudden closure which necessarily implies some novelty for him. Problem solving in a moment of insight may involve almost any kind of content. The idea may be artistic, mechanical, or theoretical. The creative idea may be administrative if it solves an organizational problem. It may be a new football play or a clever chess move or a new slogan.

If the creative act is identified by a sudden closure of insight, it is of psychological interest to consider the nature of the thinking that precedes the closure and the thinking that follows it. The thinking that precedes the closure is in the nature of worrying about the problem. In that stage the thinker may have some feeling for the nature of the solution, which is not yet verbalized. He may be able to say that a solution should look like this but not like that. He may be able to say that such and such things are associated with the solution, which is still unknown, and that certain other things seem to him to be irrelevant. The thinker may be haunted by the problem so that he deserves no credit for working at it. Because the solution has not been verbalized, it is difficult for him to discuss it, even though he may have some feeling for the nature of the solution which is to be found. It is as if the partial solution were essentially affective in nature. The solution to a problem is often so embarrassingly simple that it can be explained in a few minutes of a lecture. Instead of pride in the new idea, there might even be some hesitation in admitting the very considerable cost in time and effort. The work that follows the moment of insight may require months or years of drudgery and hard work in testing and developing the new idea.

After an idea has been found relating to the solution of a problem, it can usually be verbalized. At this stage we can begin to apply the various tools that we have been taught: the scientific methods, experimental designs, and deductive logic. The object is then to test the idea with all the rigorous procedures at our disposal. The method of attack is then quite explicit. Sometimes the solution of a problem consists of a succession of inventions, and it happens not infrequently that the principal contribution in problem solving is in the design of the experimental equipment itself. Our principal concern is not with the later deductive stages in the solution of each problem. We are interested here primarily in the psychological processes that precede each of the sudden closures at which the significant ideas appear. Most of the instruction in scientific methods is concerned with the testing of an idea, but this instruction has very little to offer in teaching students how to produce ideas. It is this pre-focal stage of the process of problem solving that specially needs investigation.

Although we seem to be unable to teach ourselves and our students how to produce ideas in problem solving, it may be worth while to consider some of the attitudes that favor the production of ideas. In teaching graduate students many of us have had occasion to observe them as well as ourselves in the work of problem solving. We shall consider a few hypotheses concerning the attitudes that seem to favor the production of fruitful ideas and other attitudes that seem to stifle them. Some students have the characteristic that when a strange problem or idea is presented, their first reaction is negative. Some very bright students seem to be able to show immediately by clear logical reasoning that the proposed idea is wrong. Sometimes the proof is so convincing that one is tempted to discard further thought about the new proposal. Even when this negative attitude is associated with high intelligence, the result is not likely to be creative. The more promising candidate is one who examines the new and queer proposal. He toys with it and speculates what the implications might be, if it could be demonstrated. Because of the very novelty of the proposal, his impulse is to wish that it could be shown to be true. He asks, in effect, how the world would be different if the proposal could be sustained. This observation leads one to wonder whether those who are inventive and creative should preferably have a certain amount of gullibility.

Some years ago Walter Dill Scott when he was at Carnegie Institute of Technology, had a test in which he asked the candidate to write some of the implications of a number of strange hypotheses. What would be the practical consequences if we had eyes both in the front and the back of our heads? He might ask the candidate to enumerate some practical consequences if we could swim more easily than we could walk. Scott's purpose was to determine the candidate's fluency of ideas. Some such procedures might be tried in relation to the present problem, but we would probably find that mere fluency of ideas does not adequately represent creative talent. Fluency in seeing implications may be an important characteristic of creative ability. Some forms of fluency may signify intelligence without implying creative talent.

Even if we cannot teach students how to produce ideas, as we can teach them how to test ideas, we might be able to accomplish something by encouraging them, as well as ourselves, to adopt attitudes that favor the production of ideas. Students often have difficulty in selecting dissertation problems. When a student asks us to suggest a dissertation problem, we might encourage him to ask a lot of questions about the field of his scientific interest. I sometimes suggest that he take as an exercise the writing of twenty dissertation proposals in the course of the following week. Sometimes the student is startled by this assignment because he has the attitude that any single idea would be welcome. Our ways of teaching are partly responsible for this type of difficulty.

It seems proper to encourage students to be critical in reading the literature, but the cultivation of such an attitude has disadvantages. When students review a paper or a monograph, they often assume that their task is to tear the paper to pieces and to make the author look foolish if it is possible to do so. Often their appraisal of a paper stops there. If a student reviews the literature with the attitude of tearing it to shreds, the resulting attitude is not wholesome for his own work. It might be better to ask him to review some papers or monographs that he is enthusiastic about. In presenting a monograph, he would then pay attention to what the author was really trying to do, what he actually did, and how far he succeeded. In criticizing a paper either because of the formulation of the problem or because of experimental and analytical errors, the student could be encouraged to suggest what he would have done with the same problem. In this way the student's work on the literature would be more constructive than if we merely remain satisfied because he has a critical attitude in reading. If courses were always concerned with ideas and what to do with them, with occasional support from the literature, the student would have less difficulty in selecting one of these ideas for his own dissertation.

One of the attitudes that seems to favor the resolution of a problem can be called the method of solving a problem by denying it. One can ask whether the formulated problem, or the half-way solution that may be available, contains some hidden premise that is implied and taken for granted but which is unnecessary for the real goal. One can try to state explicitly everything that has been assumed and which has not previously been made explicit. If any restrictive assumption or premise can be found which we are taking for granted without saying so, and if that premise or restrictive condition can be thrown out, then the solution sometimes appears. For this reason it may be a good policy not to put anything in writing, not even diagrams on scratch paper, until the central problem seems to be solved. The casual sketch may inadvertently commit further thinking to some restrictive error. I asked Professor Barnard in our mathematics department about a problem. He told me that he would think about it. A few days later he came to me with a solution, but he told me that it was trivial and that very likely I had some assumptions that were irrelevant, or else I had not told him about some restrictive conditions that I had taken for granted. The solution that he proposed satisfied the question that I had asked. He was right in his inference that I had not really stated my problem completely.

TESTING PROBLEMS

A hypothesis that should be considered in the experimental study of problem solving is that the moment of insight is often, perhaps always, in relaxed and dispersed attention. Sometimes sudden closure appears after leaving a problem in disgust, or the closure will suddenly intrude when attention is distracted.

In recent years Mrs. Thurstone and I have been working on various scoring problems. The use of tests is markedly affected by the considerable labor of scoring, especially with hand scoring. It is a practical problem to reduce the labor of hand scoring to a minimum. The use of a separate stencil for every page of every test is a complication that should be eliminated if possible. A simple solution appeared as a sudden closure in a familiar but apparently irrelevant percept. In this percept a picket fence across the street was seen through a window screen at our house. There was an optical interference effect somewhat analogous to beating in sound. In this effect the window screen was seen as a universal scoring screen and the picket fence represented a printed half-tone surface. Here the percept was affected by the problem even though the problem was not focal in consciousness at the moment of closure. Attention was dispersed and relaxed. The universal stencil does not need to be fitted to any exact location on the test paper, and the same stencil can be used for all kinds of tests. The answer spaces can be arranged in any desired patterns with the same stencil. The first thought after this closure was naturally that one should have thought of something so simple a long time ago.

I have tried one type of experiment several times with complete failure. On several occasions I have left my work on a problem to walk around the block and to give attention to other things in the hope of inducing a solution by voluntarily dispersed attention. I learned that dispersed attention in problem solving cannot be faked.

Experimental studies should be made to ascertain what kinds of mental work are done best under conditions of muscular tension with concentrated attention, and which kinds of mental work can be done best with dispersed and relaxed attention. Such studies may bear on problem solving. A related and important educational problem is to determine whether the most productive results are obtained by a driving course of instruction where all the student's time is devoted to lesson learning. Another method would be to alternate periods of intense lesson learning with periods in which the student is asked to do problem solving with the same material at a leisurely tempo and without any pressure of speed. Ordinary scholarship examinations would appraise the lesson learning but not the gains in productive abilities with the newly learned material.

We don't ordinarily hear about how professional men get their useful ideas but there is much anecdotal material on how discoveries have been made. Such stories often sound as if discoveries were quite accidental. It is more likely that the investigators had previously identified themselves with a problem in terms of which they interpreted some accidental effect. Scientific discoveries at all levels are probably not so accidental as they look to the casual observer.

When creative work is studied experimentally, serious attention should be given to the inductive factor. This is one of the most interesting of the primary factors that have been identified so far because it has been shown to transcend the nature of the content, including verbal, numerical, and spatial material. One hypothesis that should be considered is that ability in the primary factor for induction may be closely related with problem-solving ability. All of the tests that have been designed for the primary factor of induction require the subject to discover the rule or principle in the material that he is working with. It does seem that this ability should be close to, if it is not identical with, the ability for creative work. Although the inductive factor is an important component in intelligence with high saturation in the second order general factor in intelligence, it probably does not alone represent creative talent. If the inductive factor which is heavily saturated in intelligence were the same as creative ability, then there should be a higher correlation between intelligence and creative talent than has actually been found. Problem solving probably represents more than induction. Frequently it consists in reformulating the problem itself and then solving the new problem.

The scientific study of creative work will probably encounter a set of questions more or less analogous to those which marked the scientific study of intelligence. First comes to mind the question whether creative talent is some general factor which can operate in many different media, or whether there are different factors of creative intelligence in the different fields. In exploratory studies of creative work it might be well to start with the working hypothesis that creative talent is a general factor, or that it can be treated as such in first approximation. This corresponds more or less to the way in which Spearman started his analytical work on intelligence. We shall then postulate a general factor for creative work, or perhaps a group of factors which express themselves in different media. Whether a creative person becomes a poet or machine designer or composer or physicist would then depend on the various primary abilities of his intelligence.

We should consider the hypothesis that this talent is not confined entirely to the cognitive or intellective domain. It might very well happen that creative work is characterized mainly by the combination of intellective functions with certain temperamental characteristics. If this should be the case, then we should fail if we confine our efforts entirely to cognitive and intellective functions. The distinction between cognitive and temperamental traits is not so clear as used to be believed. They overlap quite definitely. An example is the first closure factor, which is well indicated by Gestalt completion tests and by mutilated words tests. The informal observation has been made that those who are especially fluent in the first closure factor are temperamentally more alert, quick, and active, and more quickly responsive than those who rate lower in the first closure factor. We cannot make this judgment about the second closure factor which has other temperamental associates. Those who excel in the second closure factor seem to be more frequently deliberate in manner. However, these are only rough observations which have not yet been checked against actual performances.

In studying normal functions it is sometimes useful to turn to the extremes of pathology. In factorial studies of the primary mental abilities we have found it useful to include in the experimental groups the most extreme differences in the abilities to be analyzed. Applying the same principle here, we should profit by gaining information about the work of geniuses. We should not confine ourselves to them because they might be very individualistic subjects that would be hard to manage. By analogy one might expect to find that ordinary people who have creative talent in some line would tend to be somewhat individualistic and this probably agrees with general observation. It may also be found that creative people are on the whole less outgoing and sociable than the average but one can think of exceptions to such a hypothesis.

Many years ago I had the privilege of working rather closely with a man who is certainly known to be a genius, and I am trying to recall in the present context some of his intellectual and personality characteristics. Immediately after receiving an engineering degree at Cornell, I went to work with Thomas Edison as one of his laboratory assistants. I talked with him daily and had good opportunity to observe his work habits. He was a man of many strong convictions and he did not seem to have much admiration for university education. It seemed to me that one of his most outstanding characteristics was a tremendous fluency of ideas. For every experimental failure he seemed to produce three more experiments to try. In this sense he seemed to be tireless. The cot in his office was probably used for lying down to think about his problems as often as it was used for sleep. Thomas Edison seemed to have a startling fluency of ideas which spread often far from the immediate problem. He seemed to have an absolutely endless array of stories. Few of them were fit for publication. Especially relevant to our present problem is the great fluency in proposing alternative solutions to a problem. This fluency of ideas should certainly be investigated in this context, but I suspect that there may be different kinds of fluency and that all of them may not be equally relevant to creative work. Another characteristic of Edison was the casual way in which he treated experimental failures. These seemed to be merely part of the day's work and the signal for starting another experiment.

In setting up experimental studies of creative work, it would seem best to arrange for two groups, namely, one experimental group of men who have demonstrated creative talent to a marked degree and a control group at the same intellectual level but with less productive or inventive talent. In order to make sure that inventive talent is really distinguished from other intellectual traits, the control group should consist of men who have high scholarship and intelligence, who are capable of profound critical appraisal in the same field as the more productive group. Both groups should be comparable in professional prestige.

One naturally considers the availability of graduate students in the universities for studies of creative talent and originality. There are serious limitations with such a proposal. In studying individual differences among graduate students in their ability to produce ideas we must necessarily depend on the judgments of their teachers. To make judgments about students as to their originality is so different from the customary academic judgments about scholarship that there is some question whether we can trust available judgments for this kind of study. We can be sure at the outset that scholarship grades will have only imperfect positive correlation with the desired ratings for originality. It should be a major part of a research program to obtain rather extensive and systematic judgments of the originality of men in various fields and to make close comparisons with the judgments of these men when they were students some years ago. Some illuminating results might be obtained from such a study which could be a part of a major research program.

One study might be made with students in musical conservatories where they could be rated on originality and promise in musical composition. Similar studies might be set up in art schools and in schools of architecture. The college curricula in engineering and in medicine are so crowded that there may not be adequate opportunity to judge the originality of students. It would be especially promising to have studies of originality in organizations such as the Bell Telephone laboratories or the Westinghouse and General Electric laboratories. In such organizations there should be good opportunity to differentiate between the originators and the men whose intellects turn more readily to critical and evaluative tasks and administrative duties. Such organizations could not survive and produce if they consisted only of innovators.

In judging men as to originality we must deal with the inevitable conflict between the standards of our day and those of the next generation. In the musical conservatories and in the art schools there might be some erratic young artists of doubtful status at the present time who will be the leaders twenty years from now. We shall certainly make errors in rating young talent. Every field of endeavor has such errors, but it is probably more common in some fields than in others. Work in the physical and biological sciences can be checked more promptly as to its potentialities than original work in the fine arts. Although every field is subject to errors of this kind they are probably less frequent in the natural sciences than in the fine arts.

It should not be necessary to wait for the results of scientific studies on creative talent before taking some action about this problem. Considerable improvement can probably be made in the selection and training of university students by giving serious consideration to creative talent. For a number of years I was chief examiner at the University of Chicago when the undergraduate college was reorganized. Scholarships were awarded annually on the basis of competitive examinations. At one time there was discussion about how the examination should be appraised for the scholarship awards. The simplest procedure was merely to add up the total number of points that the student had earned with equal weighting for the four divisions, namely, the physical and the biological sciences, the humanities, and the social sciences. This naturally gave the advantage to those students who were good lesson learners over a wide range of content. It seemed more plausible that we would select a larger number of future leaders in the various fields by following a different procedure. According to this proposal we should require only a minimum of attainment in the four fields in order to assure communication with the student at the college level. This should be the minimum requirement for eligibility. The applicant should be expected to show marked attainment in some one field. The third requirement should be some evidence that the applicant has produced something new on his own initiative. This might be some production in musical composition or in poetry or in the design and construction of scientific apparatus, some useful invention, an article or publication on some general social issue. It might even be some clever business enterprise that showed initiative and originality. The appraisal of the applicants would, of course, require the pooled efforts of university faculty groups in their respective fields. One of the objections was that these appraisals could not be reduced to machine scoring. The more serious objection was based on educational policy, namely, that we should bet on the student who could demonstrate good scholastic performance over the whole range of the college curriculum. We would probably be selecting a larger proportion of future leaders in the various fields by the policy that emphasizes independent creative effort even though it would require more faculty participation.

There have been various proposals for the selection of university students by centrally controlled examination methods. It would probably be an error to adopt any wholesale method of awarding university scholarships by means of uniform national examinations. It would be better to allow every community to select its quota of scholarship awards by various methods of selection subject to minimum admission standards that would be set by each college. The standards would vary with the communities and with the colleges.

In setting up experimental studies in this field we should certainly include objective tests of temperamental characteristics, as well as tests of the primary factors that seem to be important. These should include the two closure factors, the several space factors, the inductive factor, and several memory factors. Some of the perceptual functions may be differentiating for creative talent. Some characteristics of perseveration may be diagnostic.

One more hypothesis concerning the possible nature of creative talent and a proposed type of experimental study of it will be considered. According to one theory of intelligence (1), the psychological act originates in the essentially affective and non-verbalized, non-focal motivations and needs of the individual. The development of the psychological act towards final overt expression consists in successive particularizations. Each of these successive steps can be regarded as a choice point or bifurcation. If the choice point is in focal consciousness, then the decision is subject to rational control and it is fully conscious. If the choice point decision is made unconsciously in the particularization of the act, then the choice is determined instinctively or by intuition, habit, or chance. According to such a theory, a high intelligence is indicated by a choice point that becomes focal in consciousness at an early stage of the act so that a wide range of possible overt expressions is under conscious control. If the act develops under great pressure of motivation or emergency, then it will develop to nearly overt definition before it becomes focal in consciousness and the act is then likely to be impulsive and relatively less intelligent. We might now make an additional interpretation for this theory of intelligence. It seems plausible that there should be individual differences in the ability to be in some kind of rapport with the unconscious stages of the act, even before it becomes focal in consciousness. Imaginative people might have some kind of rapport with their own unconscious thinking before it becomes focally conscious. It might be possible for a person to have the ability for a high degree of abstraction, in that his acts are subject to conscious deliberation at their early stages, and still not be in rapport with his own unconscious thinking. Such a person would be intelligent and even profound but possibly not creative.

This hypothesis could be experimentally investigated. In one type of experiment we might start with a set of several hundred cards with a word on each card. There might be a code of several rules by which a number or a letter is associated with each word. Consider a case where only four digits are used, namely, 1, 2, 3, and 4. Man-made objects might have the even digits 2 or 4. Large objects might have the digits 1 and 2. Whatever the code is, it should refer to the meanings of the words. The subject might be asked to guess a number for each presented word. The first ten or fifteen cards would be presented with their correct digits. The next ten or fifteen cards would be presented without any digit and the subject would be asked to guess for each card. The next set of cards would have the digits given. In this way alternate sets would have the response given for each word and alternate sets would require the subject to guess. No word should ever be repeated unless it is done intentionally to give the subject the impression that he is doing a memory test. It is easy to determine for each experimental setup the probability of a correct guess. After a while the subject might say that the word "elephant" should perhaps have the number 1, that he does not really know why, and that he is merely guessing. One might tabulate the proportion of correct guesses for the successive groups of cards. When the subject sees the principle he will have all the responses correct, and the experiment is then of no further interest. Our interest would be in the proportion of correct guesses before the moment of insight. This type of study was done as a doctor's dissertation by one of my students (2). Marked individual differences were found. Some subjects continued at the probability level until suddenly the insight appeared, after which the performance was, of course, perfect. Other subjects showed a marked gain in the number of guesses up to sixty-five per cent correct before the moment of insight appeared. One hypothesis would be that those who show this gain in the guessing of right answers before insight have more imagination than those who do not show such a gain. It will be recognized that this experiment is similar to current experiments on concept formation. Instead of dealing with a general process of concept formation, we would be interested in individual differences in learning before insight. Experimental work of this sort should be leisurely with ample response time for each guess. The subject's ability to guess right more often than wrong, even when he feels that he is guessing, is the characteristic of special interest in this problem.

If a research program on problem solving is undertaken by the college board, I fear that a reorientation will be necessary as regards the criterion. The conventional criterion of college scholarship would not be adequate. Furthermore, in the exploratory studies which would be necessary to solve this problem, it is almost certain that the ordinary objective scoring methods will have to be relaxed. It is almost certain that machine scoring must be ruled out. It may eventually be possible to conduct a test for originality and creative talent with tests that can be machine scored, but it would be a serious error to impose any such restriction on research.

The recent presidential address on creative talent by Guilford(3) at the American Psychological Association is a very well-considered discussion of the problem. Some of his major considerations are the same as those we have discussed here. Guilford has proposed a rather complete factorial analysis of a large number of tests that represent various hypotheses concerning creative talent. The present paper has attempted only to outline some exploratory studies that might be concurrent with Guilford's more complete and formal program.

In this paper we have considered some working hypotheses that might be helpful in initiating the scientific study of creative talent. First, we may start with the tentative assumption that creative talent is qualitatively the same at all levels of problem solving. The creative act may be characterized by the moment of insight which is often preceded by non-verbalized prefocal thinking. The moment of insight is normally followed by explicit and deductive thinking in testing the new idea. The prefocal thinking before insight should be studied in order to gain understanding about problem solving as a form of thinking. A hypothesis has been considered that the moment of insight can be expected in dispersed attention more often than in concentrated attention. A research program on creative talent should put the main emphasis on what happens before insight. It is not enough merely to describe the curiosities of the creative product.

We might consider creative talent as determined, in a descriptive way, by the rapport that the actor has with his own preconscious thinking. This hypothesis can be considered as an extension of an earlier theory of intelligence according to which intelligence is defined by the degree of incompleteness of the act at which it can become focal in consciousness. This rapport can be studied experimentally.

The hypothesis should be considered that creative talent is in large part determined by the temperamental characteristics that are associated with intellect. It would probably be an error to look for creative talent exclusively in the cognitive or intellective domain. Even though we know little about creative thinking, we can encourage students as well as ourselves to cultivate those attitudes which favor problem solving, including tolerance for that which is novel. Experimental studies should be on two major problems, namely, to inquire about the nature of the thinking that leads to a moment of insight, and to investigate empirically how to differentiate creative talent by objective and experimental procedures. It is conceivable that we may discover how to select people with creative talent before we learn much about the nature of that kind of talent. Serious consideration might be given to a plan of teaching which alternates between intensive lesson learning and the leisurely application of the new material to original problems. Finally, let me urge a realization of the importance of this problem because the creative talent in our population is our greatest national asset.

REFERENCES

(1) Thurstone, L. L. *The Nature of Intelligence*. London: Kegan Paul, Trench, Trubner and Co., Ltd.; New York: Harcourt, Brace and Co., Inc., 1924.

(2) Bouthilet, Lorraine. "The Measurement of Intuitive Thinking." 1948. Unpublished thesis, The University of Chicago.

(3) Guilford, J. P. "Creativity," *Amer. Psychologist*, Vol. 5. 1950, Pp. 444-454.

PANEL II

Criteria for the Evaluation of Achievement Tests

Criteria for the Evaluation of Achievement Tests

FREDERICK B. DAVIS

FROM THE POINT OF VIEW OF THE TEST EDITOR

Format

THE PROCEDURES used by a test editor for evaluating an achievement test are pretty much the same as those employed by any other test technician. There may, however, be certain characteristics of tests that receive more attention from the editor than from others. Examples of these are the format, the arrangement of pages, the typography, and the texture of the paper. Studies by Tinker and Paterson (1) have suggested that conventional fonts(such as Scotch-Roman or Old Fashioned) set in 10- or 12-point type in lines about 3½ inches long are easiest to read. They recommend the use of black ink on dull-finish white paper to provide maximum legibility. It may not seem quite so unnecessary to mention the desirability of this classic color scheme if I call attention to the tinted paper and green ink used for the new *California Reading Test*, (2) which bears a 1950 copyright.

Tinker and Paterson also report that some spacing between lines is desirable; apparently, for 10-point type, the insertion of two extra leads provides adequate spacing. These specifications generally lead to the adoption of a two-column format for each page of a test, since a single 3- to 4-inch column makes an awkwardly narrow page unless a great deal of space is to be wasted in wide margins. As it happens, the nature of many items lends itself well to a double-column format. Few multiple-choice items have choices long enough to stretch across a seven-inch line. In any event, if each choice is to begin on a new line, less space is wasted if short rather than long lines are used. Sometimes, test publishers run more than one choice on a line, presumably to conserve space. Ordinarily, even more efficient utilization of space can be achieved by selecting an appropriate line length and placing the items in two or more columns. It seems reasonable to suppose that a pupil can work more rapidly and with less interference from the mechanics of finding and keeping his place if items are arranged consistently in tabular format in regular columns. Sample C for Section F of Test 2 of the new *California Reading Test*, Intermediate Form AA, is arranged as follows:

Sample C: Skating in Holland
1 Skating 2 Wrestling
3 Baseball 4 Football
5 Recreation in Holland

while item 112 is arranged:

112. Rubber Plantations in Brazil
1 Automobile Tires
2 Plantations 3 Rubber
4 Brazil 5 South America

Though the test publishers are no doubt following a consistent pattern of a sort, the effect may be to confuse a child and make it easier than it needs to be for him to make a clerical error in writing the number of the answer he thinks is correct.

The arrangement of material on the pages of a test ought to make it easy for the examinee to progress naturally from item to item and part to part without undue attention to keeping his place. Illustrations and reading material needed to answer certain items ought to be conveniently placed. A frightening example of what printers can do if they work without the supervision of a test editor is provided by the *Victory Corps Aeronautics Aptitude Test* of the U. S. Office of Education. The writer prepared the copy for this test in 1942 but had no opportunity to supervise the printing. The result was a test in which diagrams were inserted among the choices of items and on pages separate from the accompanying items so that the examinee had to leaf back and forth to answer them. Even so recent a test as Intermediate Form AA of the new *California Reading Test* presents a reading passage on page 7 with its accompanying items on the next page, which is not a facing page. The directions at the top of page 8 indicate, "You may look back to find the answers." To do so, the examinee must turn the page back and forth.

A confusing characteristic of some tests, such as the *Iowa Silent Reading Tests* (3) and the *Sangren-Woody Reading Test*, (4) is the presentation of certain items upside down so that they cannot be read without turning the page. Though intended to serve a useful purpose, arrangements of this sort introduce difficulties into test administration that can usually be avoided without extra cost if a little ingenuity is exercised.

VALIDITY

Having evaluated the format of a new test, the editor may next consider its validity, the latter being defined as the extent to which the test measures what it ought to measure. The editor may not know what it ought to measure, so he must look for evidence on this point. A rationale, with accompanying test outline, is ordinarily considered the best means of summarizing data concerning the validity of an achievement test. The rationale should state the purpose for which the test is to be used and should explain how the proposed test can be expected to serve that purpose. It should also explain how the outline was constructed. Ordinarily, the topics to be covered in an achievement test and the weight to be given to each should be determined on the basis of analyses of courses of study, textbooks that are widely used, and the judgment of authorities in the field. The test editor must undertake to evaluate the adequacy of the outline. His judgment is, of course, subjective, but gross deviations from the accepted procedures for constructing one or the absence of any outline leads him to raise preliminary questions about a test's validity.

The next step in considering a test's validity is to judge the extent to which the weights assigned to topics in the outline are actually reflected in the items. This is a complicated process and only a crude estimate is usually possible. The actual contribution of each group of items to the variance of the scores derived from the entire test can be computed for a given sample of examinees only if the variance of each group of items and the intercorrelations of scores on each group are known. Ordinarily, data of these kinds are not supplied by the test constructor and the test editor cannot easily obtain them. Hence, he may estimate the weight of each kind of item by noting the relative numbers of each type. This procedure does not always yield a reasonably accurate estimate of relative weights. For example, the total comprehension score derived from the *Diagnostic Examination of Silent Reading Abilities* (5) includes five parts with twenty items in each. Yet the contribution of part 7 to the total variance is about twice that of part 6.

The editor can hardly expect to find precise agreement between the numbers of items he assigns to the topics in the test outline and the weights specified for these topics, but marked discrepancies may suggest the need for altering the content of the test to make it measure the weighted combination of mental functions that is desired.

The nature of the individual items is the most important single element affecting the validity of a test, so the editor must pay special attention to the mental skills actually measured by each item. This cannot be done by casual inspection. The skills or information an item appears to call for are often different from those actually brought into play in answering it. Examples of this will be presented later in this article.

There seems to be an almost endless variety of defects in items. Brief descriptions of several common defects and illustrative items follow. (6)

Ambiguity of Expression.

Item writing calls for simplicity and crystal clarity of expression. Good item writers try to follow this prescription but may still use words and phrases in such a way as to result in ambiguity. Consider the following item:

It is generally agreed that the major responsibility for starting the First World War lay with
- A Germany and Austria.
- B Germany and Serbia.
- C Russia and France.
- D Russia, Austria, England and France.
- E Germany, England, Russia, and France.

The answer keyed was E. Most competent historians agree that the underlying causes of World War I lay in the motives and actions of the governments of the four nations mentioned in choice E. The item writer thought that his item would evoke this background of information, but he discovered that historians were interpreting the word "starting" more narrowly than he had. Thus, some competent historians marked choice A as correct. The item-analysis data indicated that the percentages of the high-scoring and low-scoring groups (with respect to total test score) choosing choice A were 65 and 83, respectively.

In an effort to eliminate the ambiguity of the item it was changed to read:

> Historians generally agree that major responsibility for causing the First World War lay with
>
> A Germany and Austria only.
> B Germany and Serbia only.
> C Russia and France only.
> D Russia, Austria, England and France.
> E Germany, England, Russia and France.

It has been found by experience that no one is able to detect all of the ambiguities in items that he writes. Defects can often be located by an item writer in his own items if the latter are put away for a considerable period and then restudied. Since this is usually impractical and is rarely completely effective, it follows that items should ordinarily be examined by several critics, each working independently, and then rewritten to eliminate ambiguities and other defects before experimental tryout. When item-analysis data become available, the choice-by-choice data should be examined carefully and incorrect choices that prove especially popular should be noted. Often, these data provide clues to undetected ambiguities.

Inadequacy or Incorrectness of the Keyed Answer.

The choice keyed as correct in an achievement test item should be not only the *best* of the choices presented but should be an answer that is judged to be thoroughly adequate by competent authorities. If a thoroughly adequate answer (as so defined) cannot be stated clearly and succinctly enough to fit into a usable item choice, it is usually best to discard the item. Often the same point can be tested with an item so written as to lead up to a thoroughly adequate answer that can be stated clearly and succinctly within the bounds of conventional item format. The writer must emphasize that he is not criticizing the use of the "best-answer" technique under proper conditions; he is critical only of its abuse. Failure to observe the need for making every keyed answer a thoroughly adequate answer goes far to explain why objective tests are frequently criticized by subject-matter specialists. Test constructors should not try to hide a fundamental lack of scholarship behind a smoke-screen of item-analysis data. Subject-matter specialists not only have the right to ask that each item in an achievement test conform to high standards of scholarship but they have a duty to do so. Illustrations of inadequate answers are plentiful; a couple from vocabulary tests follow:

insolent (7)
- A studious
- B envious
- C arrogant
- D accidental

27. To pour means to (8)
- 1 drive.
- 2 empty.
- 3 lack.
- 4 hurt

A person who knows the meaning of "insolent" and "pour" will not mark these items incorrectly; the keyed answer is clearly the *best* answer in each item. But the better the vocabulary of an examinee, the more likely he is to be unsatisfied with the keyed answers.

Occasionally, items that are incorrectly keyed escape immediate detection. One arithmetic reasoning item, tried out with an incorrect answer keyed as correct, proved to have a biserial correlation with total score of over .50. Four choices, including the correct answer, worked well as decoys. However, in checking the comments of subject-matter experts, it was noted that one man had keyed as correct one of the decoys. Investigation showed that most informed people were fooled by the item and got the wrong answer, but that the one critic was right. Incidents of this sort point up the necessity for having items thoroughly criticized by competent authorities before they are tried out.

Unattractiveness of Decoys.

It is desirable for several reasons to make the decoys in a multiple-choice item as attractive and as nearly *equally* attractive as possible to the examinees, (9) but great care must be exercised lest the thin line separating the *correct* or the *best* answer from the decoys be crossed. An item of the kind in which the decoys are clearly wrong, so wrong as to seem unattractive, is the following in which the examinee is asked to select the word that means the *opposite* of the boldface word:

90. **intricate** (10)
- 1 efficiency
- 2 clay
- 3 simple
- 4 delicacy

The only one of the choices that is an adjective (like the boldface word) is the correct answer. This would serve to make the other choices unattractive to examinees who consciously notice this fact or who sense it without realizing why "simple" matches up with "intricate."

A more obscure reason for the unattractiveness of certain decoys lies in their overlap. Consider this item:

Quail are usually located
- A near grain fields.
- B in swamps.
- C running across the road.
- D in marshes.

Swamps and marshes are so nearly alike that if B were an answer D would have to be; consequently, the examinee, knowing that only one answer is correct, realizes that neither B nor D can be correct. Either one alone might make an attractive decoy, but when both are used neither is likely to be effective. In this item, choice C is expressed so differently from the others and is so specific (in referring to *the* road) that it is likely to be an ineffective decoy. Hence, the examinee who knows nothing about quail and their habitats stands a good chance of marking the item correctly by virtue of logical reasoning. The item writer did not intend to produce an item measuring that mental function but his poor choice of decoys brought about that result. This illustrates how the mental skills actually measured by an item may not be those it superficially appears to measure.

More specific determiners that cause decoys to be unattractive take varied forms. The following item illustrates one of the more obvious types by which choice 4 is rendered unattractive to anyone who notices that it does not fit the "a" in the stem.

36. A bonbon is a (11)

1 heresy.
2 bonnet.
3 dainty.
4 explosive.

The logical relationships of the choices in an item may sometimes permit clever examinees to locate the keyed choice on the basis of reasoning alone, thus rendering the decoys useless for these particular examinees. Consider the following item, which has several faults:

55. The maximum number of bursts made by pilots in air combat rarely exceeds

A 3 rounds.
B 9 rounds.
C 12 rounds.
D 18 rounds.
E 25 rounds.

As stated, the item doesn't make sense. There is no point in struggling with the wording, however, because of a basic flaw that renders it useless. For purposes of discussion, however, let's revise it to read:

55. In air combat, the maximum number of rounds fired from a gun in a single burst rarely exceeds

A 3.
B 9.
C 12.
D 18.
E 25.

If choice A were correct, it is obvious that the remaining choices would have to be. If choice B were correct, choices C, D, and E would have to be. Similarly for choices C and D. Only if choice E were correct would it be logically possible to have but one correct answer. Therefore, choice E *must* be keyed if the item is to have only one correct answer. Anyone who can reason needs to know nothing about air-to-air combat to answer this item; for examinees of this kind the decoys cannot be expected to function.

A different kind of item in which some decoys are rendered unattractive because of the logical interrelationships of the choices is as follows:

> In trail bombing, the initial bomb-release line is
>
> A outside the critical zone.
> B within the critical zone.
> C the edge of the critical zone as the plane approaches the objective.
> D directly over the objective.
> E the point at which the last bomb is released.

Choices A, B, and C are such that one of them must be true. So choices D and E can be ruled out by an examinee who notices this regardless of whether he knows anything about trail bombing. Thus the item becomes a three-choice instead of a five-choice item for certain examinees. That many of the high-scoring group tested with this item may have noticed the fact that choice A, B, or C had to be the keyed answer is suggested by the item-analysis data. The percentage selecting each choice in the high-scoring group was approximately as follows: A, 13; B, 38; C, 36; D, 2; and E, 8. Three per cent omitted the item. B is the keyed answer.

The same principle operates in a more subtle way in the following item:

> An inspector is instructed to allow not more than 3% variation in cables 5/16 of an inch in diameter. He should reject a cable having a diameter of
>
> A .305 inches.
> B .313 inches.
> C .315 inches.
> D .321 inches.
> E .332 inches.

Logical considerations require that the answer to this item be either A or E if there is to be only one answer. What appears to be a five-choice item is reduced to a two-choice item on the basis of logical reasoning alone.

Double Keying of an Item.

Great care must be taken to make sure that an item has only one defensible correct answer unless specific instructions to the contrary are provided. The following item does not have a single correct answer:

> When soap and water are mixed, the combination does not separate upon being allowed to stand. Which one of the following has been formed?
>
> A A colloid
> B A solution
> C A mixture
> D A compound
> E An emulsion

According to chemists, a colloid may be referred to as a colloidal solution and also as a mixture. Hence, choices A, B, and C are all defensible as correct answers and the item is not worth trying out without replacing choices B and C.

A different type of double keying is sometimes encountered in items like number 55 reproduced in the preceding section. If the correct answer to this item—25—were made the second largest number of rounds mentioned, there would be two answers. If it were made third largest, there would be three answers, and so on. In any event, examinees well informed with respect to air-to-air firing would find more than one correct answer whereas examinees who knew nothing about the subject matter but who reasoned out the logical interrelationships of the choices would select the new choice E for reasons explained in the preceding section.

Cross-Keying of Items.

A careful check should be made to determine whether the answer to one item in a test is given away by a statement in another item. For example, one would not want to test for the spelling of a word that was used in another section of the same test. This is an obvious sort of cross-keying, but subtle examples of this type of defect are often found.

Abuse of Multiple-Choice Format.

A multiple-choice item should not be constructed merely by assembling several true-false items about a given topic. This practice is well exemplified by the following item, which is also faulty in regard to punctuation, capitalization, and English expression:

The Rook:
A Is the piece which has a round knob on the top
B Always moves sideways, parallel to sides of board, over unoccupied squares
C Is located at the corners of the board at the beginning of game
D Is useful in getting the king out of check by castling
E Don't know

The stem of this item provides no basis for selecting among homogeneous choices related to chess. In fact, some examinees might expect the choices to be related to birds. The use of the phrase "Don't know" is awkward in a multiple-choice item even though the examinees have been told to expect it as the fifth choice in each item. Except for special circumstances, it seems preferable to correct for chance success in scoring the test papers and to instruct examinees to omit an item if they don't know the answer.

Awkward and Incorrect Expression.

Item writing calls for a simple direct style devoid of ornamentation and personal idiosyncrasies. Sometimes an item can best be expressed by casting it in the form of a question. More often, the completion type is simpler, more concise, and easier to formulate. The following item on boxing represents a peculiar (and unsatisfactory) combination of the question and completion formats.

In retreating from your opponent, do you step back with your:
A Right foot going back first
B Left foot going back first
C A backward spring with both feet
D Either foot goes back first
E Don't know

Without ingenious and well-written items, no test can be maximally effective. No elaboration of statistical techniques can make up for the faults of individual items or for a lack of ingenuity and scholarship in writing or editing them. Test construction, though it necessarily requires the use of quantitative data if it is to be efficient, is essentially a creative or artistic process. It is this creative nature of test construction that makes it at once difficult and fascinating.

REFERENCES

(1) Tinker, M. A. and Paterson, D. G. *How to Make Type Readable.* New York: Harper and Bros., 1940.

(2) Tiegs, E. W. and Clark, W. W. *California Reading Test,* Intermediate Form AA. Los Angeles: California Test Bureau, 1950.

(3) Greene, H. A. and Kelley, V. H. *Iowa Silent Reading Test,* New Edition, Elementary and Advanced Tests. Yonkers: World Book Co., 1943.

(4) Sangren, P. V. and Woody, C. *Sangren-Woody Reading Test.* Yonkers: World Book Co., 1927.

(5) Dvorak, A. and Van Wagenen, M. J. *Diagnostic Examination of Silent Reading Abilities.* Minneapolis: Educational Test Bureau, 1940.

(6) For a discussion of the writing of test items, see the chapter by Robert Ebel in: Lindquist, E. F. (ed.), *Educational Measurement.* Washington: American Council on Education, 1951.

(7) Thurstone, L. L. and Thurstone, T. G. *SRA Primary Mental Abilities Test.* Chicago: Science Research Associates, 1947. Page 1.

(8) *Sangren-Woody Reading Test,* Form A, Part I.

(9) Horst, A. P. "The Chance Element in the Multiple-Choice Test Item," *J. Gen. Psychol.,* Vol. 6 (1932). Pp. 209-211.

(10) *California Reading Test,* Intermediate Form AA, Test 1, Section D.

(11) *Sangren-Woody Reading Test,* Form A, Part I.

Criteria for the Evaluation of Achievement Tests

JOSEPH J. SCHWAB

FROM THE POINT OF VIEW OF THE SUBJECT-MATTER SPECIALIST

A REVIEW of the current literature on validity in preparation for this panel disclosed the fact that there is nothing both new and true for me to say. This statement now holds a fortiori in view of the remarks of Mr. Stalnaker and those of Dr. Davis. However, the literature also showed that the truth existing in the literature is obscured and fragmented by competitive doctrinaire adhesions visible in a number of papers, and by an equally patent lack of shared experience among the various specialists who contribute to education. Failure of communication between the test constructor on the one hand and the curriculum-maker-cum-teacher on the other, was especially clear.

I have attempted, therefore to unify the disparate truths seen in the literature by supplying commonplace terms to replace specialists' terms. What emerges can be summarized in two statements. The first recognizes a practically important theoretical limit to the possibility of test evaluation. The second proposes a process and a machinery which would embody a realistic capitulation to the theoretical limits of test evaluation.

First: a test which is highly valid and at the same time highly useful is not possible in the very nature of the case. This opposition between validity and usefulness exists, if by "validity" one means a high degree of conformity to a certifiable standard on the part of a test. This opposition arises because a certifiable standard simply does not exist in most of the areas of the school curriculum.

Second: education at large would benefit much more from a campaign of validation directed conjointly at the teaching process, the curriculum, the principles of curriculum construction, and tests, than it now profits from a campaign which treats only the test as a variable in the system of educational factors.

The absence or inaccessibility of a certifiable standard for validation of tests may be illustrated first by the medical curriculum. Sub specie aeternitatis, the medical curriculum is a correct, complete, and effective instrument for the creation of successful medical practitioners; its principles, the powers, habits, and bodies of knowledge possessed by successful practitioners of medicine. To create such a curriculum in fact would require, first, identification of successful practitioners of medicine; second, correct and complete discovery of the powers, habits, and bodies of knowledge which make the practitioner successful; third, determination of the learning "experiences" which would inculcate or evoke these powers, habits and bodies of knowledge; fourth, determination of ways in which these "experiences" could be instituted; fifth, development of facility in the staff in the application of such "ways." If such steps as these were, in fact, the history of the medical curriculum, a valid test would be a possible and useful goal of a test-making organization. If such steps as these were even possible, there would still be point in the expenditure of funds and energies for the production of valid tests.

But in fact, the medical curriculum neither has been, nor can be, produced by any such valid route. There undoubtedly is such a thing as "successful" medical practice, but it defies clear delineation, and is itself the subject of endless debate. Should it be definitively arrayed by some miracle, the problem of identifying its practitioners remains; and this too would be a matter of doubt and opinion, since success in medical practice is difficult to distinguish from the mere reputation or appearance of success. Should this improbability be achieved, there would remain the problem of identifying the habits, powers and bodies of knowledge which distinguish the successful from the less successful. And even were this solved, the problems of curriculum construction and execution would remain. Only if all these problems were already solved, would it be prudent and proper to expend funds and energies on the single-minded validation of tests alone.

Meanwhile, what we have is a medical curriculum continually subject to change, as its professors wrestle unsystematically, and as best they can, with the welter of opinion concerning the nature of medicine, the preconditions for its achievement, and the means for evoking and inculcating these means in students. At all these points, "validation" is the need; examiners and test-construction organizations would do much greater service to education were they to consider ways in which tests, and statistical studies of test results could be made to serve as foci for the systematization of this larger problem of validation, than they now perform by their expensive pursuit of the isolated fantasy of the "valid" test.

What is true of the ancient and established profession of medicine is true, a fortiori, for general education where the curriculum is of more recent origin, and has an end in view much less concrete than a medical practitioner. This end of general education has something to do with the profession of being a man and a citizen. What a man and a citizen is, or ought to be, however, is subject to far more dispute and elicits far more radically different answers than the question of the nature of a good medical practitioner. Hence, it is obvious enough that a "valid" test of success of general education in general, which would have to proceed by measuring the extent to which each graduate was a good man and a good citizen, is a palpable absurdity.

If the examiner capitulates to this limitation and seeks instead to measure the success of a particular general education curriculum by measuring the extent to which its products possess the abilities and knowledge in terms of which this particular curriculum has defined man and citizen, the case is little better. The College of the University of Chicago, for instance, at present conceives its contribution to adequate humanity and citizenship to be largely intellectual, and to consist in aiding its students to become able and judicious in their employment of knowledge and their enjoyment of artistic productions. If we were complacently sure that this is the whole and proper aim of a general education, a narrowly valid test of this capicity to use and enjoy would be a useful instrument. But we are not; hence a more valuable kind of test would be one whose items would serve to suggest to the teaching staff alternative or additional aims of education which deserve consideration. And still more valuable would be a test which would disclose the unknown and unanticipated consequences (in addition to those intended) which any effective curriculum inevitably must produce upon its students. Only a test of this kind would be truly "valid" as a measure of the effectiveness of the general curriculum in advancing its students in the professions of humanity and citizenship, and a test of this kind would be explicitly and purposely "non-valid" in the ordinary sense of "validity."

Even if employment of knowledge and enjoyment of art be taken as an adequate statement of the whole aim of general education a problematic situation still exists which invalidates the notion of a useful and "valid" test. The problematic situation arises out of the indeterminacy of such matters as: *what* knowledge is most useful to the student; the nature of *judiciousness* in its use; the constituent habits and abilities which constitute it; the nature of the object of art; the abilities and bodies of knowledge constituent of the capacity to enjoy. These problems are subject to continuous discussion, debate, research and *revision*. Therefore to embark today upon the task of constructing a valid standardized test based upon today's tentative and merely working definition of the principles of the curriculum is to produce a test instrument at some future time, when, in all likelihood, the matters tested would have been transformed beyond relevance to the "valid" test so produced. And it would be merely petulant for examiners to demand a moratorium on improvement of the curriculum so that they might carry out the ritual dance of "validation."

What is true of the concept of education's aim at Chicago is also true of the concept of education's aim held elsewhere. The concept itself is incomplete and tentative. The terms which constitute it are only partly understood and are continuing subjects of investigation. The terms, in their shadowy and incomplete condition, are in turn represented by curricular intentions and programs whose appropriateness to the terms as ends is a matter of doubt and of continuing enquiry. Finally, the classroom practices which are supposed to realize curricular programs and intentions are valid only to the extent permitted by the competence, the experience, and the state of health possessed by the teacher at the moments when he teaches. A valid test in the ordinary sense is nothing more than an effort to measure the contours of a shifting and altering shadow; the standard for validation does not exist.

The secondary and primary curriculum contains its own version of this guessiness, this flux and uncertainty. The teaching of the arithmetic operation of multiplication is not an end in itself, and the most perfect test of the students' ability to perform this operation is of small value. The operation of multiplication assumes some meaning and significance when it becomes a necessary part of a larger problem, such as is represented by the "story problem" of the arithmetic book. However, a test which measures (however perfectly) the students' ability to translate statements about bananas or working hours still is of limited significance, since ultimately, the capacity to perform the operation of multiplication takes on its full significance when it is competently performed by a person who detects the proper need for it in a real situation complete with plausible irrelevances, concomitant matters, and affective as well as rational distractors. A test of mere "face validity," which measures the students' capacity to perform arithmetic operations or solve a "story" problem is the less useful as it is the more valid in the ordinary sense, since the test which impresses, in virtue of its trappings of validity coefficients and standardized achievement levels, is sedative instead of stimulative to the teacher's search for better means of instruction and more adequate conceptions of his aims.

Put backwards, this set of relations is more poignant. The real purpose of arithmetic teaching in the common school is to give its students the ability to perform the appropriate operations of arithmetic upon appropriate matters properly selected by him from the congeries of factors (rational and affective) which are presented to his attention by real situations. This real aim is represented in the school by the thin, grey shadow of the "story" problem. And this thin grey shadow is in turn often (and properly) represented by the thinner, greyer shadow of the operations themselves, divested even of the paper trappings of the "story" problem.

A test which speciously boasts of face validity in virtue of its one-to-one correspondence with only the operations of arithmetic, which are in turn embraced as standards of validity for their plausible-seeming "concreteness" (which is in fact, merely the simplicity of meaninglessness which comes from being out of context) is really only a shadow of a shadow of a shadow. The delusion that it is real, shared from the examiner to the teacher, a delusion enhanced by the trappings of validation and certification constitutes the sense in which "valid" tests may be harmful rather than useful to education as a whole. Far more desirable would be tests and usages of statistical data which kept the teacher alert to the incomplete and merely representational character of the moment-to-moment content of a curriculum, and which aided and stimulated his effort to put incomplete parts into relation with one another, to make symbols more vividly representative, and to search out and invent means for replacing shadows by the things themselves.

The flux and uncertainty which characterize education arise from two related human limitations, and a realization of what they are and that they are can help the examiner in his effort to make his work most useful to education. One of these limitations is the human incapacity to apprehend complex and abstract wholes completely or entirely correctly. As statistical research workers, for example, we are often mistaken in the effective conception of science which guides our choice of problems and methods. As examiners, we are unclear as to the meaning of "validity." As curriculum makers, we are more often wrong or incomplete than right about the nature of our subject matter: for example, we may have inadequate ideas about the nature of artistic works, about what historians can and cannot know, or about the literal character of scientific conclusions and theories.

In the case of general education, the whole which concerns us has something to do with a good and happy life for the men our students will become. And whatever other errors we may make when we come to actualizing this abstract notion in a curriculum, the first and commonest error is to be mistaken or incomplete about the abstract idea, itself. Our conception of it may arise from conscious thought or be only the product of unexamined experience; we may even be unaware of the operation of an ethical postulate in our curriculum planning. It functions, however, and it is more often inadequate than correct. And however precise and sound may be our subsequent acts of specifying this abstract notion, choosing materials of study, determining sequence and relation in the curriculum, deciding upon teaching methods and techniques, establishing calendars and time-allotments to parts of our program, the results can be no better than our originating conception of a good and happy life. For "good and happy life," one may read "physician," or "engineer" or "statesman"; the fact that we are always only partly right about these abstract wholes remains.

The question of the nature of these significant wholes, these ends of education, is one perennially settled and perennially re-opened. Hence, a conventionally valid test of the ends of general, professional or common education, is impossible. Yet an important function of tests, vis-à-vis these wholes, remains. That function is to hold up before the curriculum maker an informing image of his own view of these ends and of the views of them held by others, For abstract conceptions are corrected and made more complete only as men with diverse views of them confront intelligible statements of each other's views and engage in discussion and debate concerning them. The examination has its place in this process because of the difficulty of communicating abstract ideas. Exposition of the meaning we assign them often fails because to hold one view of a given idea is also often to have only one set of terms with which to describe it, and the terms of description may be as strange as the view to be described, to one who holds a different view. We therefore have recourse to examples: if we cannot adequately state the ultimate ends of our curriculum, we hold up the curriculum itself in the hope that through it the observer or colleague can understand the ends it aims at. But a curriculum is only incompletely embodied in any single concrete object. A list of lovingly selected, revised, and polished units of instruction fails to reflect it because such a list omits the many alternatives which were rejected in its making, and gives little hint of the debate, study and discussion out of which meaningful rejection and selection arose. The same can be said of instruction itself: how many of us would wish our purposes in education to be represented in an observer's mind by what he is likely to grasp from a chance visit or two to our classroom?

In the last analysis, the adequate communication which is the only means for improving our conception of ends of education, can arise only from a great variety of embodiments of these ends. With a variety of embodiments or reflections, what is erroneously or incompletely suggested by one is corrected or completed by another, in much the same way that an induction of the characteristics of a class of objects is brought about by correcting the judgment formed from a few members of the class by submitting it to the effect of examining other and varying members of the class. The test or examination for a program of education is a special kind of embodiment of the ends of that program: it is concrete where essays about ends are abstract; its unity can be grasped in a single study; it tends to be precise at just those points where curriculum descriptions are ambiguous; it is the product of leisured thought and correction where acts of instruction are necessarily the result of the instant's decision. It is because of these virtues of tests as a communicating medium, and because of the need for such communication in the improvement of the curriculum that it is important that test-construction agencies consider ways in which tests, their statistical study, and other resources uniquely available through such agencies, can be used not as mere "valid," and therefore static measures of a static curriculum, but as centers and foci for the discussion and improvement of all parts of the curriculum, including tests.

The second kind of intellectual limitation is not so much mistaking the whole as it is the failure to recognize and correct the consequences of making analyses of that whole which rightly or wrongly we take as the purpose of our educational program. Thus, if we propose to teach students judicious use of the fruits of scientific inquiry, a shattering analysis is required before the purpose can be translated into a curriculum which has real existence from day to day and month to month. We must determine that scientific inquiry consists of this and this and this component act, with this and this effect upon the knowledge which accrues from it. Then we must move to an analysis of "judicious use," deciding that it involves this and this and this activity. We must further proceed to an analysis which introduces yet another gap between the whole in its original form, and its representation by parts of the curriculum: by hopeful guesses, we must join some notion of the learning organism to our analysis of what is to be learned, in order to say that this and this and this activity, which constitute the act of judicious use, are to be developed by these and these and these study and classroom activities.

There are two ways in which such an analytic process can go astray. In the first place, there is the inherent weakness of analysis per se: that no fracture of a complex and interacting whole but fails to lose the whole. This is as true for the ends of a curriculum as it is of a living organism. For instance, a student taught on the one hand to "generalize data," and on the other to correct an incorrect inference, or make a correct one, and so on through a list of alleged components of "scientific method," may become competent at each of these tasks taken alone, but remain incompetent to solve a problem in which several of them are required in some particular relationship. The situation is not unlike the relation between arithmetic operations and the "story problem."

In addition to the inevitable shadow of unreality which defaces the curriculum as a consequence of the act of analysis itself, there is the defacement which arises from particular errors and failures of particular analyses. For it is inescapable that a given principle of analysis shall overemphasize differences between the parts it discriminates and obsure differences which would appear were another analytic principle employed. For instance, a program in the humanities concerned with literature may analyze its mass of material in terms of *genres* with the result that what is unique and important about the lyric poem, the tragedy, the novel, or historical writing, will become clear and significant to the student. But what is unique and important to all *genres* in virtue of their origin in a given place or epoch will necessarily be obscured. Conversely, a course founded on an historical analysis into periods and places will clarify what is unique and important about the mass of literature of a given time or place, but obscure important differences in *genres*.

Inadequacies which arise from analytic failure are to be corrected by the same means required for correction of mistakes about wholes. In order to perceive that his parts of the curriculum do not reconstitute the whole with which he started, the instructor must confront evidence that students exhibit the partial capacities which constitute his curriculum yet do not possess the whole which he intended to give them. In order to perceive that his particular analysis of his subject matter emphasizes some aspects of it at the expense of others, he must be given the chance to contrast what his students can do with what students trained in differently analyzed curricula can do.

It is the need for these contrasts and comparisons which again illustrates the mutually exclusive character of "validity" and "usefulness." The more valid the test for a particular curriculum, in the sense of more perfect conformity to it, the less its usefulness in pointing toward inadequacies in, and illuminating possibilities of correction and completion of, that curriculum. It is these same needs for contrast and comparison, however, which constitute a task and challenge to the examiner: (1) to produce tests which are not mere slavish replicas of a given established curriculum, but which go beyond the instructor's apparent intentions toward his real ones, so that he may become aware of his opportunities for enlarging and correcting his curriculum; (2) to effect, through a variety of tests as foci, communication and discussion among instructors representing varied interpretations of common ends, so that they be reminded and informed of the limitations of their isolated constructions. Nor is the examiner's protest that he is not his brother's keeper a valid reason for eschewing this responsibility.

The invalidity, in the case of the examiner, of Cain's rejoinder to God, can be suggested by indicating the sense in which central-agency examination construction as currently organized pretends that test construction is merely an engineering problem. The essence of the engineer outlook is the notion of separation of ends and means. Someone else tells the engineer what to do. His job is merely to do it. In the case of education, a Dean, Chairman, or Committee states the content and objectives of a curriculum. The examiner receives the statement, and makes a test to fit it, testing its goodness of fit by statistical analyses supplemented by minimal review by the originating dean, chairman or committee. This is a convenient and pleasant way of life for the examiner, insofar as evasion of responsibility contributes to pleasantness. And it can speciously be defended on the ground of efficient division of labor, a ground whose plausibility is enhanced by appealing to the specialized knowledge demanded by the examiner's task. But efficiency in the accomplishment of a task is not properly measured by the dispatch with which a single component of the job is done; for one component of a complex machine may be produced with dispatch only at the price of reducing the effectiveness of the whole machine of which the component is to be a part. This is happening in the present "efficient" separation of examination-construction from the remainder of education; the virtual obsolescence of almost any elaborately standardized and validated test is prevented from being obvious only because the separation of test and curriculum, of examiner and curriculum planner and teacher, hinders the communication which would reveal it.

There is one other way in which an essentially invalid character accrues to "valid" tests in virtue of non-communication among participants in the educational process. It arises from the fact that a conventionally "valid" test may literally test something other than it intends to test, and fail to be caught in this act by the usual statistical methods of validation. This possibility stems from the infinite ingenuity of the student in devising ways to move from a question to an answer. An analogy can be found in biological research where there is need to measure some profoundly buried quantitative factor in the life process. In virtue of its inaccessibilty, no attempt is made to measure it directly. Rather, an "indicator" of it is sought, an indicator which can be expected to vary as the wanted factor varies, as, for instance, red blood cell frequency or hemoglobin concentration, as indicators of the capacity of the body to supply its parts with oxygen. Then some method is devised for estimating the quantity of the "indicator." It is easy to become aware of invalid links between the method of estimating the indicator, and the actual fluctuations of the indicator, for these factors are literally on the surface, intrusively in the field of sense experience, as, for instance, the common way of measuring blood pressure by measuring the pressure necessary to compress a vessel carrying the blood.

But the possibility that the "indicator" is invalidly taken as a measure of the prime factor is neither obvious nor easy to determine. Yet the possibility is a *prima facie* probability in every case. For it is extremely unlikely that the value at any given moment of any variable quantity in an organism, is a function of one other factor only. The same is clearly true of examinations bearing upon any intellectual skill. The question put, and the alternatives posed for the students' judgment, may be such that the answer can be arrived at by exercise of the skill or ability which the item is intended to measure. What remains problematic is the *actual* skill, ability, or other route, by which the individual studies does in fact arrive at the proper response. The obvious instances of this all of us know: the possibility of a student's determining the right response in an information test by using his wit vis-à-vis the way an examiner's mind works; a student's circumventing the intentions of an information item by estimating logical probabilities in the situation; circumvention of a test of the power of inference by drawing upon information. But whether routes alternative to the one intended do or do not exist in a wide variety of tests which we currently accept as valid tests of particular skills, can best be determined only by drawing the student, as well as the examiner and instructor, into the conversation which should obtain among educators. Tests need to be tested by "depth" interviews in which the student is taught to think aloud as he answers test items, and thus provide some evidence as to the modes of intellectual operation employed to answer them.

The broader conception of validity outlined thus far suggests that test-construction agencies work toward two relatively neglected aims. First, to assist the individual educator, concerned with his own, indigenous program, to bring his principles of curriculum construction, the curriculum itself, and his tests to a better and more real approximation to his own ultimate ends; second, to bring educators concerned with the same specialty but operating at different institutions into communication with each other so that their conceptions of the ultimate ends themselves, as well as of means, will, by exchange of ideas and of convictions, better approximate the true and useful.

Pursuit of these aims would require four changes in existing organizational structures and procedures of test-construction agencies, relating respectively to their personnel, the tests themselves, the analysis of tests, and the use to which tests and test-analyses are put.

1. Concerning personnel, it is proposed first that test analysis become the special responsibility of a small group of specialists designated for the function. Second, the actual construction of tests and items be the responsibility neither of test specialists, recently graduated Ph.D's., nor ad hoc committees superficially supervising the work of one of the aforesaid. Rather, test construction should be the responsibility of individuals who represent in their persons the experience of working with or toward more than one curriculum in his special field and level of education. Such persons would be designated for their examining function by being sought among the staffs of institutions engaged in the development and execution of programs in the field to be tested. Such a person would be released during a stated period for a third or a half of his time to the test-construction agency. His task would be to construct not one, but several kinds of tests, each of which would be a formulation *in concreto* of what he thinks his program in science, or humanities, or history, is trying to do.

2. The intention of such a personnel structure is the production of two kinds of *varieties* of tests. One would be a variety of embodiments in several distinctly different tests (not alternative forms of a similar kind) of what is purportedly the same curriculum; this variety, elicited from a representative of one curriculum is to function in the improved approximation of that curriculum to its ultimate aims. Second, in each field and level, at least three persons would be sought, the three to be representative of as widely differing conceptions of education in that field as seems feasible. The variety of tests which emerge from such varied representation of the field is intended to function to bring educators of widely differing backgrounds and convictions concerning the field into communication with each other for the purpose of sharpening, broadening and rectifying their ideas and convictions concerning it.

3. This double variety of tests becomes maximally useful for its designed purpose only if submitted to appropriate statistical analysis. Such analysis would include conventional validation procedures, but these procedures would be aimed at the broader goal. Correlations among the test units constituting the set of tests of a single curriculum will be designed to exhibit its unity or lack of it. Lack of it may be interpretable either as an unexpected complexity in the ends of the curriculum, or of ambiguity in the conception of these ends held by the examiner and his staff-colleagues. The presence of it may be interpretable as an unexpected simplicity or naiveté in the curriculum as conceived. In either case, what is sought is initiation of

thought concerning the tests and the curriculum, of experiment with their methods, materials and procedures, and ultimately, considered change in the curriculum and tests.

Correlational and other procedures imposed upon different test-sets representing different curriculums and upon single representatives of several test-sets will be designed to determine the degree of real difference in their effects upon students of apparently different curriculums. It should be noted that in this situation there are two groups of students for each est-set under consideration: students trained in the curriculum represented by a given test-set, and students trained in other curriculums. The possibilities inherent in the series of contrasts thus afforded are too numerous and complex for analysis here. The end of such analysis is, however, simple. It is to bring into the vivid meaningfulness afforded by contrast what it is that each participant curriculum does and does not do for its students. The ultimate aim is the same as before: to initiate thought, experiment, and improvement of the participating curriculums.

In addition to analysis of the relations inter se of various images of a single curriculum, and of the images of various curriculums, it is further proposed that ordinary validating procedures be supplemented by "depth interviews" intended to elicit from students in the act of taking tests, the content and mode of thought employed for the purpose. Experiments to determine whether such procedures are possible and fruitful have been initiated and appear to speak affirmatively.

4. The potentialities inherent in tests so constructed, designed and analyzed can become real only if the tests and their results are made genuinely accessible to their consumers, the participant educational institutions. For this purpose, numerous conferences and workshops are means, and responsibility for their initiation and administration becomes a routine function of the test-construction agency.

Parallel to the two ends of the testing program, the two varieties of tests, and the two areas of analysis, the types of conference would also be dual. One would be concerned with explication of the results of analysis of various tests of a single institution's program, and with speculation as to the significance and cause of the discovered results. It would take place among the teaching staff concerned with the tests and program in question, the staff member responsible for the tests, and the test specialist concerned with their analysis. Out of such conferences should grow a more informed teaching staff, more meaningful instruction, as well as concrete experiments with tests and the curriculum.

The second type of conference or workshop would be concerned with the tests and analyses representative of a variety of curriculums. It would be concerned with the significance to each curriculum of the others it saw reflected in the variety of tests, and of the results exhibited by the measurements of achievement by its own students and those of other institutions, on the variety of matters tested. It would take place among at least two representatives of the several curriculums involved, together with the staff members who constructed the tests and the specialists who supervised their analysis. Out of such workshops should grow a new and larger order of thought, experiment, and revision of curriculum than is remotely possible in the present insular state of most educationists and most educational institutions.

Criteria for the Evaluation of Achievement Tests

JOHN B. CARROLL

FROM THE POINT OF VIEW OF THEIR INTERNAL STATISTICS

One of the most interesting questions to ask about an achievement test is the question of its *homogeneity*. When we ask this question, we may be concerned with whether all the items of the test measure the same ability, *or* we may be concerned with whether a test has sampled the information contained in a homogeneous field. For example, we may be asked whether an American history test contains a representative sample of the body of facts contained in a given syllabus of a course in American history, and whether the items uniformly measure achievement in learning this particular syllabus.

Unfortunately, most discussions of test homogeneity thus far have dealt with ability tests, that is, with tests which are intended to measure the level of difficulty which can be mastered by the examinee. Now, some achievement tests are intended to place the examinee on a continuum of task difficulty—e.g., a vocabulary test, but a large number of achievement tests are *not* so intended. Rather, they are designed to measure the examinee's total knowledge of an area. The score is merely an expression of the probability that the subject can pass any item—we don't necessarily care which items he passes. A valid measure can be obtained even if the items do not vary in difficulty (item difficulty being measured by the proportion passing each item). How can we measure the homogeneity of such an achievement test?

I will return to this question later, but let us review some proposals which have been made regarding the homogeneity of ability tests. We see at once that there are several different idas of what is meant by homogeneity. One technique of determining homogeneity is the factor analysis of the items of a test. From the standpoint of factor analysis, if the common factor variance of the items is traceable to only one factor, the test would be said to be homogeneous. Notice this, however: unreliability of the items would not affect one's conclusion regarding the homogeneity of the test, so long as the item intercorrelations constituted a matrix of rank one. If the matrix were of higher rank, the technique would presumably enable one to sort out the items into two or more homogeneous sets if the data so permitted. Nevertheless, the study of the homogeneity of test items by factor analysis meets several obstacles. Aside from the rather exhaustive labor required, there is the question of what kinds of correlation coefficients should be factored. On this question, I have advocated the use of tetrachoric correlation coefficients, corrected, if necessary, for the effect of chance success. In a paper (1) read at the last APA meeting, I attempted to demonstrate techniques for determining the factorial purity of an ability test. The Seashore Sense of Pitch Test was my example; it was found to be about as factorially pure as one might desire. The items acted homogeneously in placing the examinee on a single difficulty continuum. Nevertheless, my techniques are complicated and require a number of assumptions.

A somewhat different approach is adopted by Jane Loevinger in her monograph, "A Systematic Approach to the Construction and Evaluation of Tests of Ability." (2) First, she is careful to point out that she restricts her discussion to "a specific class of tests: power tests of ability." She defines a perfectly homogeneous test of an ability as one "such that, if A's score is greater than B's score, then A has more of some ability than B, and it is the same ability for all individuals A and B who may be selected." From this she deduces that when the items of a homogeneous test "are arranged in order of increasing difficulty, every individual will pass all items up to a certain point and fail all subsequent items." She proposes an index of homogeneity based on the variance of scores relative to the variances of perfectly homogeneous and heterogeneous tests. Now, Loevinger's concept of a homogeneous test implies that the test measures a *difficulty continuum*. It is analogous to Guttman's idea of a perfectly reproducible scale. It must be pointed out that unreliability of items from any cause will depress Loevinger's index. For example, the index cannot have a value of unity unless all items are perfectly discriminating. Consequently, it is likely to be misleading. I applied Loevinger's index to my data on the Seashore pitch discrimination test and found a value of only .08. Loevinger's index can be corrected for some sources of unreliability. For the Seashore Test data, an estimate of the homogeneity corrected for the effect of chance success is .42, and this still does not correct for the effect of an imperfect operating characteristic. (I must omit discussion of the correction techniques because that is not my present concern.) The point still remains that Loevinger's homogeneity is the homogeneity of a difficulty continuum, and it is depressed by unreliability due to any source.

We do not yet have available a simple and effective index of the homogeneity of an ability test. Nevertheless, let us now turn to the problem of measuring the homogeneity of achievement tests. It is obvious that either the factor analysis approach or Loevinger's approach will yield a decision of test heterogeneity when applied to many types of achievement tests, particularly those whose items do not vary greatly in difficulty. It must be that a different kind of homogeneity is implied here.

I would like you to imagine a hypothetical model of a test which might be regarded as a homogeneous achievement test. Suppose we have a number of individuals who have been given an opportunity to learn a number of facts. At the time of testing for knowledge of these facts, each individual is conceived to have a true score represented by the probability, p_i, that he will pass any item. The value of p_i varies from individual to individual, but whatever its value, it applies uniformly to all items. In the sense that the probabilities are uniform over the items, the test is said to be homogeneous. Starting from any series of true individual p_i values, it is possible to work out the expected variance of observed individual scores and the variance of item difficulties. I developed a hypothetical series of data for this case, using random numbers to decide whether my hypothetical individuals passed the items. Now, I would call this test perfectly homogeneous, but Loevinger's index of homogeneity was only .32. Even this figure is spurious, because the items vary in difficulty only by chance; Loevinger's index tends to be biased positively because of chance variations in item difficulties.

We can also think of the case where for any individual the probabilities of passing the items vary in a systematic manner. For example, the probability of passing any one item might be always proportional, or otherwise functionally related, to the probability of passing any other item. We can still call the test homogeneous because of the systematic variation of the individual-item probabilities. This type of model takes care of the homogeneous achievement test whose items vary in difficulty.

I have some data on a test which seems to exemplify this model. The test is Thurstone's Word-Number Memory test, and while it would ordinarily be regarded as an ability test (of the memory factor) it is, as it were, an achievement test in a nutshell, because it tests the subject's knowledge of 20 word-number associations to which he has just been exposed in a learning period. The total score is the number of items recalled, and from the many factorial studies of memory tests, this score measures the rote learning factor in a highly valid manner. In order to see whether the items of the test are functioning homogeneously in the production of the total score, I wanted to find whether the p_{ij}'s were systematically related. By p_{ij} I mean the proportion of persons at the i^{th} ability level passing item j (whose difficulty, p_j, is measured as the proportion of the total group passing the item). To get reasonably accurate measures of p_{ij}, we often have to combine individuals in a sufficient range of ability to build up the number of cases. I tabulated the p_{ij}'s for the 20 items at 6 ability levels, and thus obtained a 6 x 20 matrix. Arranging the rows and columns in order of the marginal proportions, it was evident that some law governed the entries. In fact, it turned out that when the logarithms of the p_{ij}'s were taken, the resulting matrix had a rank very nearly equal to 1. The rows and columns, in other words, were proportional. Furthermore, it seemed that the following equation were satisfied, at least within small errors:

$$\log p_{ij} = W \log p_i U \log p_j,$$

where W and U are proportionality constants, and p_i and p_j are the marginal proportions. (I don't know whether this equation can be rationalized; at the moment, it is purely empirical.) Thus, knowing an individual's ability level and the "difficulty" of the item for the whole group, it was possible to estimate quite accurately the probability that an individual of that ability level would pass the item. The prediction was so close that the correlation between the predicted and obtained p_{ij}'s was .95. Even this correlation is not as high as it might be because of the sampling error that we would expect for the obtained proportions. In any case, I propose to say that this result indicates that the items of the test are homogeneous.

Thus, whenever we can predict, to a high degree of accuracy, the probability that an individual of ability level i will pass an item of difficulty level j, the test will be called homogeneous. A possible measure of homogeneity is the correlation between the obtained p_{ij}'s and the predicted p_{ij}'s, adjusted for the expected sampling variance of the obtained p_{ij}'s.

Now, the interesting thing is that this formulation can be extended to ability tests measuring a difficulty continuum. Take any test, whether you think it's an ability test or an achievement test, and make a three dimensional plot, using these variables as coordinates:

a) The ability level of the individual, based on total test score;

b) The over-all difficulty of each item; and

c) The proportion of individuals at a given ability level passing an item of given difficulty.

If the resulting swarm of points is systematic—e.g., following a slanted plane, or a curved surface—and if the variation of the points from the fitted surface is not beyond what one might expect by chance variation, the items of the test will be said to be homogeneous. Furthermore, items which show significant deviations of their p_{ij}'s from the fitted surface will be indicated as non-homogeneous with the generality of the items.

There remains the problem of the fitting of surfaces to the three dimensional swarm of points. I have not had enough experience with this method to give you any discussion of this beyond two statements: a function of the form

$$\log p_{ij} = W \log p_i U \log p_j$$

is likely to be satisfactory and convenient for achievement tests. For an ability test, the normal ogive will perhaps generate the best fitting surface.

To sum up, I am proposing that homogeneity be defined as a property of a test such that the probability that an individual of a given ability level will pass an item of a given difficulty level is uniformly predictable within expected sampling errors for all items and all ability levels. In this sense, homogeneity is *not* the same as the property which concerns the function of a test in placing individuals along a difficulty continuum. A perfect measure of a difficulty continuum is necessarily homogeneous by the foregoing definition. But an achievement test whose items do not vary in difficulty could also be perfectly homogeneous by this definition.

REFERENCES

(1) "Problems in the Factor Analysis of Tests of Varying Difficulty." *American Psychologist*, Vol. 5, 1950. Pp. 369. (Abstract)

(2) Loevinger, Jane. "A Systematic Approach to the Construction and Evaluation of Tests of Ability. *Psychol. Monog.* Vol. 61, No. 285, 1947.

Criteria for the Evaluation of Achievement Tests

HAROLD GULLIKSEN

FROM THE POINT OF VIEW OF THEIR EXTERNAL STATISTICAL RELATIONSHIPS

LET US consider the relationship of an achievement test to various possible types of external criteria. It seems to me than an achievement test may be evaluated first in terms of its relationship to a total battery of aptitude tests. Secondly, it may be evaluated in terms of its relationship to training, practice, experience, or drill in the subject concerned. And third, it may be evaluated in relationship to future performance of the individual in various related tasks.

With respect to the [illegible] method of evaluating the relationship of the achievement test to a battery of aptitude tests, we may point to several concrete illustrations of this type of study. We usually think of the aptitude test standing or falling in terms of its relation to a criterion. I propose here that the criterion should stand or fall in terms of its relation to aptitude tests. For example, in studying the relationship of the Basic Navy Battery to grades in service schools it was found that the U. S. Navy Reading Test correlated well with grades in Gunner's Mates Schools. Steps were taken to introduce more practical work on assembling and disassembling the guns and to place less emphasis on a study of manuals in the training and testing of gunner's mates. (3, 4) Such a change in the criterion resulted in a decreased validity for the Reading Test and an increased validity for some of the mechanical type tests. Grades in Torpedoman School correlated high with Reading and the General Classification Test, and low with mechanical knowledge—an inappropriate pattern for Torpedoman grades. An increase in the validity of the Mechanical Knowledge Test and a drop in the validity of the Reading and General Classification Tests followed the introduction of achievement tests in the Torpedoman School (7) (page 308).

Another concrete illustration of the use of a set of validity coefficients for making an initial appraisal of a criterion is furnished by the Navy Basic Engineering School (7) (pp. 305-309). The first eye-opener to an unsatisfactory situation was the fact that an arithmetical reasoning test gave the highest correlation with final grades in the Basic Engineering School. Mechanical knowledge and mechanical comprehension tests gave the lowest correlation with these final grades. This state of affairs was extremely peculiar in view of the fact that arithmetic constituted *one* seventh, and shop work *four* sevenths of the curriculum. Investigation of grades showed that the standard deviation of the grades covering the one week of arithmetic was about 8.0; the standard deviation of grades for approximately four weeks devoted to shop work was about 2.5. Averaging one seventh arithmetic (standard deviation 8.0) with four times as much shop work (standard deviation 2.5) gave a final grade which correlated in the high .80's with arithmetic and in the .50's with shop work. It, of course, would have been possible to suggest that grades in shop work be multiplied by a factor of 100 or so to increase their weight. However, a study among different instructors showed that the ratings of shop work intercorrelated from —.11 to .55. Clearly the problem was to develop more precise ratings before allowing them to determine the major portion of the grade. This was done in terms of better rating scales, various gauges and calipers to measure accurately the precision of the product.[a] The result was a highly reliable set of ratings of shop performance which were judged by the chiefs to be ratings of important characteristics. A recheck on the correlation of various aptitude tests with grades after institution of this program showed that the mechanical knowledge tests had the highest correlations and the arithmetical reasoning tests the lowest correlations with final grades[7] (p. 307). Here we have a case where the interrelationships among part scores, the inspection of the curriculum, the judgment of the instructors, and the pattern of validity coefficients from selection tests all gave a consistent picture. The initial picture was strongly indicative of a *low intrinsic validity* of basic engineering grades. In the final picture all the evidence pointed toward a *higher intrinsic content validity* of basic engineering grades.

Comprehensive Factor Study of Criterion and Predictor Variables. Extending the idea of the previous section we see that the most information about the criterion would be afforded if a comprehensive matrix of intercorrelations including both criterion and predictor variables were available. A representative set of predictor variables should be included so that the investigator will be able to inspect both the high *and low* validity coefficients to be certain that this pattern is reasonable. Also a variety of modes for assessing different aspects of the criterion must be used in order to determine the dimensionality of the criterion variables and the extent to which the various expert judgments agree and disagree.

---

[a] The major part of this work was done by Dr. Nicholas A. Fattu on NDRC Project N-106 of the Applied Psychology Panel. This Project was conducted by the College Entrance Examination Board at the request of the Bureau of Naval Personnel.

As yet there are very few factor studies which include an attempt both to analyze and represent various *aspects* of the criterion and to include a representative list of a variety of predictor variables. (1, 2, 5, 6)

It seems to me that such studies are the major means now at the psychologist's disposal for evaluating what may be called the intrinsic validity of a criterion. They would enable us to see the extent to which ideas regarding the nature of the criterion and the nature of the predictors were correct. For example, a differential validity study has found that the best predictor of grades in Latin is the clerical aptitude test. (8) This fact might be considered relevant to the view that Latin constitutes a valuable mental discipline in clear thinking.

Let us now consider possibilities of evaluating achievement tests in terms of their relationship to some type of training or practice in the material. Here we would need to utilize parallel achievement tests, one given before and one after the training or practice. Clearly the experts' judgment regarding the content of these tests is corroborated if the students *do not* know the material before training and *do* know it afterwards. However, if the material is known before training or is not known after training the expert judgment would bear reconsideration. I do not have any rules to suggest as to amount of improvement that should be expected as a result of training—whether good students should improve to the same extent or a greater or lesser extent than poor students. However, the behavior of the material with reference to instruction or practice seems to me to be a matter that must be investigated carefully in the case of every achievement test.

The third type of evaluation suggested is in terms of future performance on related tasks. Here again we must depend primarily on the judgments of experts in the field as to the expected relationship between the test and subsequent performances. For example, if instruction in high school English is supposed to improve a person's ability to write personal letters or to do adequate work in English in college then the relationship between various achievement tests in English and agreed-upon desirable future outcomes could be studied. The studies in each case would depend it seems to me primarily on the type of claims that were made for the achievement test in question. For example, if Latin teachers seriously maintain that one studies Latin because he thereby becomes better in English it is relevant to study the relationship between future performance in English for students who have had and who have not had Latin, for those who have done well in Latin and poorly. If the English teacher does not make this claim and if no one else makes it then such a study is not relevant for evaluation of achievement tests in Latin.

Concluding, then, I would feel that considerably more attention should be paid to some of the external relationships of achievement tests; their relationships to subsequent relevant achievement; their relationship to training, practice, or drill in the field; and that achievement tests may be evaluated in terms of their relationship to batteries of aptitude tests.

REFERENCES

(1) Carroll, J. B. "The Factorial Representation of Mental Ability and Academic Achievement," *Educ. & Psychol. Meas.*, Vol. 3, (1943). Pp. 307-332.

(2) Comrey, A. L. "A Factorial Study of Achievement in West Point Courses," *J. Educ. Psych. Meas.*, Vol. 9, (1949). Pp. 193-209.

(3) Frederiksen, Norman. *Statistical Study of the Achievement Testing Program in Gunner's Mates Schools.* (Navpers 18079), (1948). Pp. iv + 41.

(4) Frederiksen, Norman and Monroe, Lt. A. E. *The Development of Achievement Tests for Gunner's Mates Schools.* (OSRD Report No. 5259) (Project N-106, Report No. 17), (1945). Pp. vii + 253.

(5) French, J. W. "The Description of Aptitude and Achievement Tests in Terms of Rotated Factors." *Psychometric Monograph No. 5*, 1951.

(6) Sisk, H. L. "A Multiple Factor Analysis of Mental Abilities in the Freshman Engineering Curriculum." *J. Psychol.*, Vol. 9, (1939). Pp 165-177.

(7) Stuit, D. B. *Personnel Research and Test Development in the Bureau of Naval Personnel.* Princeton N. J.: Princeton University Press, (1947). Pp. xxiv + 513.

(8) The Psychological Corporation. *Differential Aptitude Tests Manual.* Pp. E-14.

DISCUSSION

PARTICIPANTS:

Robert L. Ebel, Frederick B. Davis, Joseph J. Schwab, Harold Gulliksen, Robert L. Thorndike, Emma Spaney, John C. Flanagan, John B. Carroll, Eugene D. Carstater, Warren G. Findley, Donald A. Peterson.

Chairman Thorndike: The afternoon's papers are now before you for discussion. Let me first ask whether any of the members of the panel would like to specifically comment upon things which other members of the panel have said. Then perhaps we can throw it to you folks on the floor to raise questions and make comments as you see fit on the presentations of the afternoon.

Dr. Ebel: I would like to comment on Dr. Davis's paper. I am in hearty agreement with 99 per cent of it, and am eager for the publication of some of those sample items. The point with which I would like to take issue is the statement that the keyed answer to a multiple-choice question should not only be the best answer, but the thoroughly correct answer. I would interpret "thoroughly correct" to mean an answer so correct that no one can imagine any possible improvement of it. If that is the case, the questions we can ask in achievement tests would be seriously restricted.

I took the trouble to jot down a few questions which might be important questions in a multiple-choice achievement test. Take a test of contemporary affairs, for example. One might ask, "Why did President Truman commit American forces to combat in Korea?" To get a manageable item you can't insist on a thoroughly correct answer. Likewise, it would follow from Dr. Davis's requirement that all of the vocabulary-type items which are based on synonyms or substitutes would be ruled out, because each word, according to the dictionary, is unique in its function; you can't exactly duplicate it with any other word.

I think the second difficulty with this "thoroughly correct answer" criterion is that it would drive us to emphasize types of questions which are quite undesirable. "What does Professor So-and-So say was the cause of World War II on page so-and-so of his text?" I object very strongly to such questions. We have used them far too much in the past.

Now it is possible that I have misinterpreted what Dr. Davis meant by "thoroughly correct." There is another question, which I am afraid I was responsible for originally, which asks, "What is the position of the moon at high tide?" I used as the correct answer, "The moon is at the zenith." Those of you who live in this territory know that such an answer isn't correct at all. Even though it happened to be the best answer among those given, it should not be used as the correct answer, since one can easily express a much closer approximation to the actual facts.

I am not in favor of needless inaccuracy or carelessness in expression, but I do think that the best-answer type of multiple-choice question is infinitely superior to the "thoroughly-correct answer" type.

Dr. Davis: No, I wasn't misunderstood; I meant exactly what I said. I just think you have to exercise a little bit of ingenuity and modesty in asking questions. Now, for example, I wouldn't ask why Mr. Truman committed American forces to action in Korea. I would have it read, "What was one important reason why Mr. Truman committed American forces to action in Korea?" and then you could give an absolutely correct answer. You might have one of the many in the complex of reasons that motivated him. In other words, the first thing that you have to do is to limit yourself with respect to the types of stems that you make up.

Now let's take another example of the kind where we say, "Professor Sidney Fay has said that the major cause for the First World War was . . ." I agree with Dr. Ebel; I don't think that you ordinarily want to limit an item to one man's opinion; but rather to say: "It is generally agreed that . . ." or "Competent historians generally agree that . . ." and then within the limits of the agreement of subject-matter specialists about that you can get an answer which is absolutely correct. The second general point I should like to make is that the law of diminishing returns sets in when correct answers to items are prepared. My impression is that it doesn't set in very early in the process; and I am willing to go quite a long way up the curve to get a very precise answer, or else discard the question or change the stem. But I don't suppose that you would want to insist on such minute, detailed accuracy that building test items becomes an impossibly pedantic job.

Dr. Schwab: I would like to speak to that too, if I may. I would like to suggest the *reason* for preferring best-answer items in which the best answer is also a nearly correct one. The reason is that test instruments, willy-nilly are teaching instruments, which suggest to students the kinds of answers to give to questions. I should also like to generalize this matter widely by giving the name for it which Dr. Davis refrained from giving. The name is vulgarization of subject matter in two respects. One respect is illustrated by the history example: it is the vulgarization which pretends that the tentative and the hypothetical is the known. We see it in items in which dogmatic assertions of causes and reasons are demanded where they cannot be known as in the case of why Mr. Truman sent troops to Korea or what is the cause of the Second World War. These are only hypotheses to historians.

The second sort of vulgarization seen in many tests consists of believing that the truth is very simple and requires no qualification. Dr. Davis commented on the brevity of alternatives and the margin that exists on either side of them as an indicater of the lack of qualification in best-answer items and a failure to use to the maximum the possibility of best-answer items which are all alike in their main clause, but differ in respect of their qualifying, subordinate clauses.

Dr. Gulliksen: I think I disagree. It seems to me that we don't know enough about item writing yet to say that the discrimination among various degrees of badness must be excluded from examinations.

One of the types of items that we utilized in the social science examination at Chicago consisted of a question and several different answers. Sometimes it was a question that we had given in previous examinations—as a free-answer question on which the students wrote a paragraph. Then, instead of having the students each write very hastily under the examination conditions their own answer and having the instructor sweat to decide which was better, we reversed the procedure and let the students grade the various answers. We tried, frequently, to present five possible answers in order to have items of greater difficulty. As I recall, we always failed when we demanded complete staff agreement on the relative merits of the five choices. But we could frequently find three answers such that a fair proportion of the staff would agree that, while a given answer wasn't perfect, it was better than either of the other two; and although another answer wasn't completely wrong, it was worse than either of the other two. Item analysis data showed that these items correlated fairly well with scores on the total six-hour examination. Of course, it wouldn't have disturbed me if the correlation had been low, because it was possible that the type of discrimination required for these items was not represented in the rest of the examination.

I am surprised that Dr. Schwab should feel that students need to be protected so asiduously from developing erroneous concepts. His fear that utilizing answers which are "best" but not "perfect" will constitute bad training seems to me to be unwarranted. Surely the world is an awful place to let the student get into at all, if you are really concerned about such influences of examinations. Having such items in the examination may be very good. Maybe those items aren't as good as others, I don't know, and I doubt if anyone else knows. I would just like to speak up for keeping in items which require the student to discriminate among various *degrees* of correctness or incorrectness.

Dr. Schwab: No disagreement on that.

Chairman Thorndike: Would somebody from the audience like to introduce another theme?

Dr. Spaney: I should like to ask a question about this common sense that we should exercise in the validation of tests. I should like to know whether we might use as guideposts for this common sense some kind of ultimate criterion, such as examination of the end product, and, if so, whether we could have any helps in systematizing that analysis of the end product.

Chairman Thorndike: Does anybody want to help Dr. Spaney?

Dr. Flanagan: I should like to record a point on that. It would take too long to go very far into it, but the talk of shadows and images has disturbed me and I should like very much to see us get back to facts and substance and data. It seems to me that we have an approach to going out into the life situation and gathering data in the form of critical requirements, the things that make the difference between success and failure in the practical situation, and base our tests and our criteria on these. It is a long story, but I feel that it should be mentioned that there is an alternative to the shadows and images which have been suggested.

Dr. Schwab: I should like to speak to both remarks. Dr. Flanagan illustrates very clearly one of the doctrinaire adhesions to which I made passing reference in my talk. One axis of doctrinaire adhesion consisted of a line of which one extreme consisted of persons who felt that they deserved the name "no-nonsense" people. The no-nonsense people turn out to be simply people who have honed a problem down until it looks simple. Their "common sense" view of reality looks good because it is an unexamined notion of what reality is.

It is good *un*common sense to conclude that if a reading test correlates better than a machine-building test with scores in a course intended to teach people to build machines there is something wrong with the curriculum or with the machine-building test. But one isn't granted such easy situations very often in the primary, secondary, and general educational level. What you are supposed to be teaching there cannot be pinned down to such simple facts as whether the student can put a machine together or not, so the matter of ends requires an awful lot of arguing back and forth and clearing up one set of no-nonsense against the no-nonsense of many other people. What is required is conversation.

Dr. Gulliksen: I would just like to supplement that by saying I think any conversation which is promoted would probably be more valuable if you took care to couch it in such terms that the participants did determine whether they agree or disagree with each other.

Dr. Schwab: That in itself, of course, is the crux of every conversation, and once that be achieved, the conversation is almost over. Harold Gulliksen knows this very well. The imposition on a group of people of a ready-made set of terms is going to warp seven-eighths of the participants. The phrase "coming to terms," I think, is a phrase of importance here, and one charged with much more meaning than we usually give it. It is the problem of conversation: to try to come to terms; to begin to understand one another as the necessary basis for determining where agreement and disagreement lie.

Dr. Flanagan: I call this conversing you are talking about by teachers who don't know much about the facts of life a pooling of ignorance.

Dr. Schwab: But that is precisely the way in which all research is done, isn't it?

Dr. Flanagan: That isn't the way I do it.

Dr. Schwab: Then I really fear for your results, because it seems to me the first condition to discovery is recognition of ignorance and the delusion of knowledge, believing that one knows, for instance, what deduction or induction is, or that science is certain, is the firmly padlocked door to any reinvestigation of the question of what science is.

Chairman Thorndike: But if we wait until we *do* know to start making our tests, perhaps it will be our great, great grandchildren who will make them rather than ourselves.

That is, the doctrine of infinite regress which you were expounding to us earlier, may be an attractive doctrine, but it is for us here today a doctrine of futility. One need not be on one or the other side of this problem; one may recognize the need to examine one's postulates and assumptions at the same time that one goes ahead and does what one can at the present level of one's ignorance.

Dr. Schwab: That is what I proposed. I proposed that tests constructed by curriculum makers and teachers put through the mill of validation are one of the best foci on which to start the conversation. By no means did I postpone the test. If I think there is an infinite regress, there is no point in postponing, because you postpone forever. I did not present infinite regress as a doctrine for determining practice. I stated it as posing the problem, and I merely said that the problem required not merely creation of instruments which fitted some arbitrarily chosen standard, but rather interaction between the tested tests and the tested standards with an aim to the improvement of both.

Dr. Carroll: Don't you mean, instead of infinite regress, successive approximation?

Dr. Schwab: That is much the more complimentary term.

Dr. Carstater: There might be some point in referring to the program of this session to see if there is some integrating influence. The problem here is to consider the potential contribution of various sub-criteria for the evaluation of achievement tests. A part of this problem is identification of available criterion measures and assessment of their suitability. We had a panel of four men who looked at four different types of criteria—in the sense of evaluative procedures. Out of these "looks" we get some question as to whether any of these types of criteria, in and of themselves, have any degree of validity as criteria.

Is face validity (the editor's contribution) a criterion against which you can tell whether the measurements obtained from a test will do what you want them to do? Will examination of the degree and nature of subject-matter orientation of the test content lead to indices of the validity of the test? Is homogeneity within a test a measure of the value of that test as a measure of whatever kind of achievement you want to assess? What kinds, if any, of external relationships with test results can be depended upon to indicate whether a measure is valid for what we want to use it for?

We come to an impasse, an exercise in futility, because we have no sure measures, no sure guides, no ultimate measured criteria. So we have to come back to "common sense" and rely finally on some kind of informed judgment, perhaps best expressed in the pooled guess.

Dr. Schwab: I like very much the phrase "pooled guess."

Chairman Thorndike: Would somebody else like to make a contribution?

Dr. Findley: Before we get too far away from the illustration that was raised by Bob Ebel here and answered by Fred Davis and Dr. Schwab, I would like to make a point—perhaps it is a minor, technical point, but I think it has some significance. The question was raised as to the appropriateness of the question: "Why did President Truman commit the American forces in Korea?" Fred Davis's response was that the way to straighten that matter out was to turn it around and say: "One of the reasons why was this," and then have that more completely or perhaps completely correct. That solution is probably impractical in many cases and is probably not the best teaching approach, to take Dr. Schwab's point of view.

I think that perhaps in that situation many of us would turn the matter around and say, "Which of the following did not influence Mr. Truman in his choice?" which leads us into one of the types of ambiguity that Fred Davis might have mentioned at the beginning, but perhaps is a minor one. Perhaps the thing we want to do, since there were doubtless a number of factors in the situation, would be to turn the matter around completely and say that: "All of the following factors influenced Mr. Truman's decision except . . ." and give a series of, say, five statements, four of which were doubtless factors in the situation, one of which was clearly not—although not so obviously out of character as to make the question ridiculously easy—and in that way we would be making a question that had a clearly correct answer that was not ambiguous in its format, and that was good from the teaching point of view that Dr. Schwab wanted to emphasize.

Dr. Gulliksen: In case you could get whoever was approving this test to agree, I would certainly advocate considering something like saying: "The most important of the reasons why . . ."

I think one of the values of the objective test is that it does put the conversation in a form where people can tell whether they agree or disagree with each other. That is why I think Dr. Schwab's point is quite important, that you tie the testing very closely to the teaching.

It may be good to see if the students can discriminate degrees of importance, provided the judgment to be made is clearly indicated.

Dr. Carstater: May I comment a bit on that particular thing also? The pitfalls of objective testing in social studies are so many. When you talk about the most important reason in anything that has a political implication, it is not a question of importance, nct a question of significance, but what reason will appeal most to the most people.

Dr. Gulliksen: Reword the stem, then, if that is the way you like it.

(**Speaker**): Since you came back to this point, I would like to raise the issue Dr. Ebel and Dr. Davis got into originally. I think the question of having a true answer and a best possible answer is the relation between testing facts for themselves and testing the application of facts, and I don't see how, on the application of facts, you can get anything else but the best possible answer, rather than the true answer as such.

Dr. Davis: Your stem asks for that; it asks the examinee to make a judgment as to which of stated choices is considered best, and you can state by what criterion the judgment is to be made. Such an item might be hard to key. The possibilities in writing items are endless. My only point is: Whatever you do, be sure that the answer that you key is pretty close to being absolutely right for the one particular question that you have asked. I think you can approximate that quite well.

Dr. Carstater: To push my other point just a bit further, you can't key that as a best answer or correct answer until you have counted all the responses, and then the one most of them pick is the right answer.

Dr. Davis: There are various ways of doing it. That is one. You might want to say, however, that it is generally agreed by historians that such-and-such is the case. Nobody knows whether it is *true*; historians may be wrong, but if that is what they have agreed on, you can ask whether the person who is taking the test knows that is what they agreed on.

Dr. Carstater: But who decided on the degree of agreement?

Chairman Thorndike: But who decided on who is the historian?

Mr. Peterson: There is considerable indication that test editors tend to throw out responses to multiple-choice items where the student shows too much ingenuity in arriving at the correct answer, even though he lacks knowledge about the particular item. It seems to me the ingenuity is interesting in and of itself. I wonder if anyone has tried this type of approach: Instead of throwing away item responses in the situation that you consider bad, save those responses and put them to the stem of a question which has no answer, and see to what extent the students eliminate those that they ordinarily eliminate when they don't know the answer to an answerable question stem. It seems to me that we might be getting at a measure of ingenuity.

I presume you would have to put those questions in with a large number of other questions that were answerable so they wouldn't catch on. I wonder if anyone has attempted it.

Dr. Davis: We have worked with those kinds of problems, but I don't think we have ever tried exactly that, though I guess it would be a good idea to do so. We have had the problem, you know, on reading comprehension tests of people being better able to answer the questions based on a certain passage without the passage than with it. That has troubled me quite a bit from time to time. We have tried giving reading items to students without the passages on which the items are based, and you do measure some sort of ingenuity variable which is unquestionably of social value. I don't know of what social value.

Dr. Schwab: Mr. Chairman, if I may, I should like to claim Mr. Carstater as an ally. His comments about the nature of history and the nature of the social sciences illustrate precisely my point that if the tester does not take into consideration the curriculum and let test validities and curricular validities interact against each other and correct each other, there are some test problems that can't be solved.

Any historian, as I think this gentleman will testify, will tell you that general history is an elaborate form of fiction. Test items concerning the reasons for a political act not four years old, will always fail to receive surface validity if enough historians or political scientists are asked to look at them. And a course which teaches one of the magnificent historical fictions as if it were literal fact is a curriculum for which no tests can be valid, for the curriculum itself lacks validity.

DR. CARROLL: I should just like to comment there. Really when I talk about the homogeneity of a test I suspect I am not really measuring the homogeneity of the test but of the curriculum. I am not going to be terribly unhappy if I find that the test is unhomogeneous by my criterion; I am going to say something is wrong, possibly with this curriculum, and it is that successive approximation of throwing the ball back and forth between the test and the curriculum that perhaps provides a promised way out of this morass.

107

Appendix

PARTICIPANTS—1950 INVITATIONAL CONFERENCE ON TESTING PROBLEMS

Adams, Elizabeth C., Educational Testing Service
Adkins, Dorothy C., University of North Carolina
Allart, Shirley J., Educational Testing Service
Alt, Pauline M., Teachers College of Connecticut
Anastasi, Anne, Fordham University
Anderson, Roy N., North Carolina State College
Andrews, T. G., University of Maryland
Angell, George W., State University Teachers College (New York)
Angoff, William H., Educational Testing Service
Ayer, Frederic L., Teachers College, Columbia University

Barnes, Paul, World Book Company
Barrett, Dorothy M., Hunter College
Bartnik, Robert V., Educational Testing Service
Baxter, Brent, Prudential Insurance Company
Bear, Robert M., Dartmouth College
Beardsley, Katherine P., Barnard College
Beck, Hubert P., The City College of New York
Becker, Theodore, New York State Department of Civil Service
Benson, Arthur L., Educational Testing Service
Bergesen, B. E., Jr., Educational Testing Service
Bingham, Walter V., Washington, D. C.
Birnbaum, Owen, New York, N. Y.
Blakemore, Mrs. J. E., Board of Education, Scarsdale, N. Y.
Bloom, B. S., University of Chicago
Bolman, Ruth, American Public Health Association
Bookstaver, Renée, Teaneck, N. Y.
Braca, Susan E., Archdiocesan Vocational Service, New York, N. Y.
Brandt, Hyman, The Personnel Institute
Bransford, Thomas L., New York State Department of Civil Service
Bridges, Claude F., World Book Company
Bristow, W. H., Board of Education, New York, N. Y.
Brolyer, Cecil R., New York State Department of Civil Service
Bronk, Richard V., Teachers College, Columbia University
Brower, R. F., The Psychological Corporation
Bryan, Miriam M., Silver Burdett Company
Bryan, Ned, College Entrance Examination Board
Buckton, LaVerne, Brooklyn College
Buckingham, Guy E., Allegheny College
Burke, Paul J., American Educational Research Association
Burke, William W., Washington University
Burnham, Paul S., Yale University
Buros, Oscar K., Rutgers University

Campbell, Donald W., Board of Education, Newark, N. J.
Capps, Marian P., The New York City College
Carol, Bernard, Metropolitan Life Insurance Company
Carroll, John B., Harvard University
Carstater, Eugene D., Bureau of Naval Personnel, Navy Department
Cayne, Bernard S., Educational Testing Service
Chauncey, Henry, Educational Testing Service
Churchill, Ruth D., Antioch College
Clark, Willis W., California Test Bureau
Cohen, Elizabeth L., Educational Testing Service
Copeland, Herman A., Atlantic Refining Co.
Corcoran, Mary, Educational Testing Service
Couey, F., Maxwell Air Force Base, Alabama
Cowles, John T., Educational Testing Service
Craig, Robert, Teachers College, Columbia University
Crawford, J. R., University of Maine
Crissy, W. J. E., New York, N. Y.
Cutts, Norma E., New Haven State Teachers College
Cynamon, Manuel, Brooklyn College

Dailey, J. T., Lackland Air Force Base, San Antonio, Texas
Daniels, Evelyn P., Educational Testing Service
Davis, F. B., Hunter College
Davison, Hugh M., Pennsylvania State College
Deemer, Col. W. L., Headquarters U. S. Air Force, Washington, D. C.
Derrick, Clarence, Educational Testing Service
Detchen, Lily, Pennsylvania College for Women
Diederich, Paul B., Educational Testing Service
Diers, Helen, Teachers College, Columbia University
Dion, Robert, California Test Bureau
Doppelt, Jerome E., The Psychological Corporation
Dragositz, Anna M., Educational Testing Service
Dressel, Paul L., Michigan State College
Du Bois, Philip H., Washington University
Duran, June C., California Test Bureau
Dyer, Henry S., Harvard University

Ebel, Robert L., State University of Iowa
Engelhart, Max D., Chicago City Junior College
Epstein, Bertram, The City College of New York

Fan, Chung-Teh, Educational Testing Service
Fattu, Nicholas A., Human Resources Center, Champaign, Illinois
Feinberg, Robert M., University of Florida
Feldt, Leonard S., Rutgers University
Ferguson, Leonard W., Metropolitan Life Insurance Company
Findley, Warren G., Educational Testing Service
Finkle, Robert B., Metropolitan Life Insurance Company
Fischer, John H., Teachers College, Columbia University
Flanagan, John C., American Institute for Research
Flemming, Edwin G., Burton Bigelow Organization
Franklin, Jean, Educational Testing Service
Freeman, Paul M., Maxwell Air Force Base, Alabama
French, J. W., Educational Testing Service
Frederiksen, Norman, Educational Testing Service
Forlano, George, Board of Education, Brooklyn, N. Y.
Furst, Edward J., University of Michigan

Gallagher, Henrietta L., Educational Testing Service
Gardner, Eric F., Syracuse University
Gehlmann, Frederick, Science Research Associates
Gerberich, J. R., University of Connecticut
Gilmore, John V., Boston University
Gordon, Hans C., Board of Education, Philadelphia, Pennsylvania
Greco, Marcell P., Teachers College of Connecticut
Green, Dorothy, City College of New York
Gregg, George W., Maxwell Air Force Base, Alabama
Green, Bert F., Jr., Educational Testing Service
Gulliksen, Harold, Educational Testing Service

Haggerty, Helen R., Personnel Research Section (AGO)
Hagman, E. R., Greenwich Public Schools
Hart, Rita V., Standard Oil Company of New Jersey
Hausrath, A. H., Educational Testing Service
Healy, Ernest A., Jr., Center for Psychological Service, Washington, D. C.
Hedlund, Paul A., New York State Education Department
Heston, J. C., DePauw University
Hieronymus, A. N., State University of Iowa
Hoberman, S., Municipal Civil Service Commission, New York, N. Y.
Hoffman, E. Lee, Educational Testing Service
Hogan, Ralph M., Office of Naval Research
Hogen, Elizabeth, Teachers College, Columbia University
Horton, Clark W., Dartmouth College
Hosman, Major Richard S., Maxwell Air Force Base, Alabama
Huddleston, Edith M., Educational Testing Service
Hunt, Thelma, George Washington University
Hurd, Charles P., Columbia University
Hyatt, E. J., Cortland (N. Y.) State Teachers College

Jacobs, Audrey, The Personnel Institute
Jacobs, Robert, Educational Records Bureau
Johnson, A. Pemberton, Educational Testing Service

Jones, Edward S., University of Buffalo
Jones, Vernon, Clark University
Judkins, Nancy C., Wilson Junior High School, Mt. Vernon, N. Y.

Karlin, John E., Bell Telephone Laboratories
Keller, Robert J., University of Minnesota
Kinney, Lucien, Stanford University
Kirkpatrick, Forrest H., Bethany College
Klare, George R., The Psychological Corporation
Knotter, Margaret C., Educational Testing Service
Kogan, Leonard S., Institute of Welfare Research
Kolkebeck, Robert F., Educational Testing Service
Krohn, Earl H., Educational Research Corporation
Kruglov, Lorraine, The City College of New York
Kurtz, Albert K., Pennsylvania State College
Kushner, Rose E., The City College of New York

Langmuir, C. R., Syracuse University
Lannholm, G. V., Educational Testing Service
Layton, W. L., University of Minnesota
Leach, Kent W., University of Michigan
Leach, Sarah C., The Psychological Corporation
Lennon, Roger T., World Book Company
Lev, Joseph, New York State Department of Civil Service
Leverett, Hollis M., American Optical Company
Levy, Beatrice, Educational Testing Service
Lifset, M. A., New York State Department of Education
Lincoln, A. L., The Lawrenceville School
Lindquist, E. F., State University of Iowa
Litterick, W. S., Stephens College
Lord, Frederic, Educational Testing Service
Lorge, Irving, Teachers College, Columbia University
Lorr, Maurice, Veterans Administration, Washington, D. C.
Lurie, Walter A., National Community Relations Advisory Council

McCall, W. C., University of South Carolina
McCann, Forbes E., New York State Department of Civil Service
McCarthy, Peggy, Educational Records Bureau
McCord, Mary, University of Chicago
McGillicuddy, Marjorie, New York State Department of Civil Service
McGuire, M. Anne, New York State Department of Civil Service
McLaughlin, Kenneth F., U. S. Naval Academy
McNamara, W. J., International Business Machines Corporation
McNemar, Quinn, Stanford University
McQuitty, John V., University of Florida
MacPhail, A. H., Brown University
Manuel, H. T., University of Texas
Martin, Lycia O., State Teachers College, Trenton, N. J.
Mayans, Frank, Jr., Teachers College, Columbia University
Mazza, Marguerite, Greenwich Public Schools
Mollenkopf, W. G. Educational Testing Service
Moseley, Nicholas, Sandy Hook, Connecticut
Muhyi, I., Teachers College, Columbia University
Myers, Charles T., Educational Testing Service
Myers, R. C., Educational Testing Service

Nill, Kathryn F., Rinehart and Company

Olsen, Marjorie, Educational Testing Service
Orleans, Beatrice, Bolling Air Force Base, Washington, D. C.
Orleans, Jacob S., Board of Higher Education, New York, N. Y.
Orleans, Joseph B., George Washington High School, New York, N. Y.
Orshansky, Bernice, Teachers College, Columbia University

Pace, C. Robert, Syracuse University
Pashalian, Siroon, New York University
Peebles, Clarence M., Northwestern University Dental School
Perlman, Mildred, New York City Civil Service Commission
Perry, W. D., University of North Carolina
Peterson, Donald A., Life Insurance Agency Management Association
Plumlee, Lynnette B., Educational Testing Service
Potts, Edith Margaret, The Psychological Corporation
Prescott, George A., World Book Company

Raskin, Evelyn, Brooklyn College
Reid, John W., Albright College
Remmers, H. H., Purdue University
Ricciuti, H., Educational Testing Service

Richardson, M. W., Richardson, Bellows, Henry & Co.
Ricks, J. H., The Psychological Corporation
Riegel, Eleanor J., Educational Testing Service
Rimalover, Jack K., Educational Testing Service
Rivlin, Harry N., Queens College
Robbins, Irving, Queens College
Roca, Pablo, Department of Education of Puerto Rico
Rock, R. T., Fordham University
Roessle, R. L. B., Standard Oil Company of New Jersey
Rogers, Miles S., Educational Testing Service
Rulon, Phillip J., Harvard University

Scates, Douglas E., Queens College
Schaefer, Willis C., University of Maryland
Schrader, William B., Educational Testing Service
Schroedel, E. C., International Business Machines Corporation
Schultz, Douglas, Educational Testing Service
Schwab, Joseph J., University of Chicago
Seashore, Harold, The Psychological Corporation
Segel, David, Office of Education, Washington, D. C.
Selover, Robert, Prudential Insurance Company
Shaplin, Judson T., Harvard Graduate School of Education
Sharp, Catherine G., Educational Testing Service
Shaycoft, Marion F., National League of Nursing Education
Sichel, H. S., National Institute of Personnel Research, Johannesburg, South Africa
Smith, Allan B., University of Connecticut
Spaney, Emma, Queens College
Spaulding, Geraldine, Educational Records Bureau
Speer, George S., Illinois Institute of Technology
Spence, Donald, Teachers College, Columbia University
Spence, Ralph B., Teachers College, Columbia University
Spencer, Lyle M., Science Research Associates
Spooner, George A., Teachers College of Connecticut
Stalnaker, J. M., Educational Testing Service and American Association of Medical Colleges
Stanley, Julian C., Jr., George Peabody College for Teachers
Stauffer, A. J., Columbia University
Stenquist, John L., Department of Education, Baltimore, Maryland
Stewart, Naomi, The American Jewish Committee
Stokes, T. M., Metropolitan Life Insurance Company
Sullivan, R. H., Educational Testing Service
Swineford, Frances, Educational Testing Service
Symonds, Percival M., Teachers College, Columbia University

Taber, V., New York State Education Department
Terral, Joseph E., Educational Testing Service
Thorndike, Robert L., Teachers College, Columbia University
Thurstone, L. L., University of Chicago
Thurstone, Thelma, Chicago Public Schools
Tiedeman, David V., Harvard University
Tinkelman, Sherman, New York State Education Department
Todd, Charles B., Mitchell Air Force Base, Long Island, N. Y.
Torgerson, W. S., Educational Testing Service
Travers, Robert, Board of Higher Education, New York, N. Y.
Tremer, Carl E., New York State Department of Civil Service
Triggs, Frances, Committee on Diagnostic Reading Tests, Inc.
Troyer, Maurice E., Syracuse University
Traxler, Arthur E., Educational Records Bureau
Tucker, Ledyard R, Educational Testing Service
Turnbull, Mary E., Educational Testing Service
Turnbull, William W., Educational Testing Service

Vaughn, K. W., Hyde Park on Hudson, N. Y.
Voss, H. A., Special Devices Center, Port Washington, N. Y.

Wadell, Blandena C., World Book Company

TESTING PROBLEMS

WALKER, John, Mr. Vernon (N. Y.) Public Schools
WANTMAN, M. J., University of Rochester
WATSON, Walter S., Cooper Union
WECHSLER, David, New York University
WEITZ, Henry, Duke University
WENZEL, Bernice M., Barnard College
WESMAN, A. G., The Psychological Corporation
WHITNEY, Alfred G., Life Insurance Agency Management Association
WHITMORE, Howard S., New York Telephone Company
WILKINSON, Richard, Southwest Missouri State College
WILKS, S. S., Princeton University and Educational Testing Service
WILSON, John T., Office of Naval Research
WINGO, Alfred L., Virginia State Department of Education
WOLFLE, Dael, Commission on Human Resources, Washington, D. C.
WOLMAN, Benjamin, City College of New York
WOOD, Ray G., Ohio State Department of Education
WRIGHTSTONE, J. W., Board of Education of the City of New York

ZUBIN, Joseph, Columbia University